Nancy-Université
ATILF / CNRS
44, avenue de la Libération
B.P. 30687
54063 Nancy, France



Académie universitaire
Wallonie-Europe
Université de Liège



Modélisation d'un discours étymologique

Prolégomènes à l'informatisation du *Französisches Etymologisches Wörterbuch*

Thèse présentée par
Pascale Renders
en vue de l'obtention du grade de
Docteur en Langues et Lettres
sous la direction de
Mary-Guy Boutier et Éva Buchi

Année académique 2010–2011

Nancy-Université
ATILF / CNRS
44, avenue de la Libération
B.P. 30687
54063 Nancy, France



Académie universitaire
Wallonie-Europe
Université de Liège



Modélisation d'un discours étymologique

Prolégomènes à l'informatisation du *Französisches Etymologisches Wörterbuch*

Thèse présentée par
Pascale Renders
en vue de l'obtention du grade de
Docteur en Langues et Lettres
sous la direction de
Mary-Guy Boutier et Éva Buchi

Année académique 2010–2011

Résumé

Titre : Modélisation d'un discours étymologique. Prolégomènes à l'informatisation du *Französisches Etymologisches Wörterbuch*

Le *Französisches Etymologisches Wörterbuch*, ouvrage de référence en linguistique française et romane, est actuellement sous-exploité, en raison des difficultés de consultation que posent ses particularités lexicographiques.

La rétroconversion des 25 volumes imprimés en un dictionnaire informatisé, que la communauté scientifique appelle de ses vœux, pourrait remédier à ce problème. La densité et la complexité structurelle de l'ouvrage font toutefois craindre que l'opération se révèle peu raisonnable, voire utopique. Par ailleurs, l'informatisation présente le risque de dénaturer le discours fewien et d'ouvrir la voie à des pratiques de consultation incorrectes.

Cette thèse se propose d'étudier la faisabilité du projet de rétroconversion du FEW, en prenant en compte toutes les contraintes qui lui sont imposées.

Dans la première partie de l'étude, nous modélisons le discours étymologique fewien de façon à résoudre les difficultés de consultation et de lecture relevées, tout en respectant les structures de l'ouvrage. Cette modélisation, formalisée en XML, rend compte de deux dimensions complémentaires du FEW, correspondant à deux visions de l'œuvre : comme un thesaurus d'unités lexicales d'une part, comme un recueil de monographies d'autre part.

La seconde partie de notre étude examine comment appliquer le modèle au texte fewien de façon automatisée. Un logiciel de rétroconversion a été conçu dans ce but. Le noyau du logiciel est constitué d'une trentaine d'algorithmes qui identifient, dans un article du FEW, les divers types d'information pertinents et les balisent. L'application du logiciel sur un corpus de 150 articles du FEW produit des résultats de balisage comportant très peu d'erreurs problématiques.

Le résultat est encore perfectible, mais il démontre la faisabilité d'une rétroconversion du FEW qui, moyennant la création d'outils d'exploitation appropriés (moteur et interface de recherche), devrait répondre à la majorité des attentes de la communauté scientifique, en rendant l'ouvrage plus accessible sans pour autant éluder la complexité et la profondeur de son discours.

Mots-clés : FEW, XML, informatisation, retroconversion

Abstract

Title : Modelization of an etymological discourse. Prolegomena to the computerization of the *Französisches Etymologisches Wörterbuch*

The *Französisches Etymologisches Wörterbuch*, the reference book in French and Romance linguistics, is currently underused because its lexicographic features make it hard to search and hard to read.

The retroconversion of the 25 printed volumes of the FEW into a computerized dictionary, as desired by the scientific community, may address this problem. The density and the structural complexity of the dictionary may however make retroconversion appear as an unreasonable, even utopian, endeavor. Furthermore, computerization induces the risk of skewing the fewian discourse, thus opening the way to incorrect search practices.

This doctoral dissertation proposes to study the feasibility of the retroconversion of the FEW, taking into account all of the constraints on the project.

In the first part of the study, we model the fewian etymological discourse in order to address the searchability and readability issues of the dictionary in a way that preserves its structures. This modeling, formalized in XML, takes account of two complementary dimensions of the FEW, corresponding to two visions of the dictionary: a thesaurus of lexical units on one hand, a collection of monographs on the other hand.

The second part of our study examines how to apply the model to the fewian text in an automated fashion. A retroconversion software was constructed to this end. The kernel of the software is comprised of about thirty algorithms that identify and tag, in an article of the FEW, the various relevant types of information. The application of the software to a corpus of 150 FEW articles results in a tagging that exhibits very few problematic errors.

There is still room for improving the tagging, but it already demonstrates the feasibility of a retroconversion of the FEW that, assuming the availability of appropriate exploitation tools (search engine and search interface), should address most of the expectations of the scientific community, making the dictionary more accessible without eluding the complexity and depth of the fewian discourse.

Keywords: FEW, XML, computerization, retroconversion

Remerciements

Cette thèse n'aurait pas vu le jour sans la complicité et l'enthousiasme d'Éva Buchi, directrice de recherches à l'ATILF, et de Marie-Guy Boutier, professeure à l'Université de Liège. Elles ont été pour moi des exemples d'excellence et, à tout moment de ma recherche, des guides avisés, qui ont dirigé mes travaux avec autant d'assurance que de bienveillance. Je voudrais leur exprimer ici toute ma gratitude pour la confiance qu'elles m'ont témoignée tout au long de mon doctorat.

L'initiation à la programmation et à l'algorithmique que nécessitait mon sujet de thèse m'a été procurée à l'Université de Liège et à l'Université de Louvain-la-Neuve par d'excellents maîtres, les professeurs Cédric Fairon, Pierre-Arnoul de Marneffe et Gérald Purnelle, que je voudrais remercier de tout coeur pour leur enseignement. Ils m'ont ouvert l'accès à un monde insoupçonné, que je me réjouis de découvrir davantage.

Je m'estime heureuse d'avoir pu mener cette recherche dans un environnement de qualité et profondément humain, autant à Nancy qu'à Liège. Merci aux rédacteurs du FEW, aux responsables du centre de documentation et aux membres de l'ATILF pour leur accueil chaleureux, leur amitié et leur disponibilité. Merci à l'équipe de l'ARD de l'Université de Liège pour son aide dans la recherche de financements, son soutien constant et ses conseils éclairés. Merci également aux membres fondateurs du Réseau des Doctorants de l'Université de Liège, sans lequel cette thèse n'aurait pas été réalisée en LaTeX, pour leurs initiatives visant à susciter la collaboration entre doctorants au delà des frontières facultaires.

Mes collègues du Service de linguistique du français et du Service d'algorithmique de l'ULg ont été des compagnons fidèles dans les moments de doute comme dans les moments de joie. Merci à vous, Claire, Cyril, Esther, France, Gérard, Marie, Nicolas, Thomas et Xavier, pour toutes les aventures vécues en votre compagnie. Vous avez fait de ce doctorat une aventure unique sur le plan humain autant que sur le plan intellectuel.

Parmi mes collègues, Cyril Briquet mérite particulièrement ma reconnaissance. Ses compétences en algorithmique et en programmation ont été déterminantes pour le projet. Sans lui, le logiciel de rétroconversion du FEW n'aurait pas encore vu le jour, et le balisage automatisé des articles n'aurait jamais produit de si bons résultats. Merci de tout coeur, Cyril, pour ton investissement, ta générosité et ton soutien constant, lequel a dépassé le cadre strict de notre collaboration.

Je voudrais également remercier ici mes amis, ma famille et tous mes proches pour leur amour indéfectible, qui m'a permis de tenir le coup à travers les épreuves, durant ces années parfois difficiles. Merci à François, qui a accompagné les premiers pas de cette thèse. Merci à Véronique, à Pierre et à mes parents, toujours à mes côtés. Enfin,

merci à toi, Gérard, pour l'amour et le soutien que tu m'offres chaque jour, pour tous les petits bonheurs inoubliables déjà vécus, et pour les grands bonheurs à venir.

Sommaire

Chapitre 0

Introduction

Il n'est plus besoin de souligner l'importance du FEW, ou *Französisches Etymologisches Wörterbuch*, dans la recherche sur l'histoire et l'étymologie du lexique français et, plus généralement, galloroman, voire roman. Les vingt-cinq volumes de cette œuvre monumentale, rédigée en grande partie par Walther von Wartburg, auxquels on peut ajouter aujourd'hui un grand nombre de publications annexes (constituant l'épi- ou la péristructure du FEW), représentent une référence indispensable :

> Il n'est donc pas inutile de rappeler à tous ceux qui s'occupent du lexique français que le *FEW* (ou l'*FEW*) est une somme que l'on ne peut se dispenser de consulter, même si cette consultation n'est pas toujours commode. Ce dictionnaire donne, en effet, une vue systématique de tout le vocabulaire français, à la fois dans son développement historique et dans ses variantes géographiques. Pour chaque mot, pour chaque sens, pour chaque construction, pour chaque locution, le *FEW* apporte des renseignements très précis (sous une forme elliptique que la bibliographie et surtout la pratique permettent d'élucider) [...]. (Goosse 1991, 163-164)

La prise de position d'André Goosse laisse deviner que la consultation de cet ouvrage si important est, malheureusement, loin d'être aisée. Après une investigation rapide sur laquelle nous reviendrons dans la suite de cette thèse, nous avons constaté que les utilisateurs du FEW (au nombre desquels il faut compter les rédacteurs de la refonte) étaient confrontés à quatre difficultés majeures. Un premier problème, dû à la densité du dictionnaire et à sa complexité structurelle, est l'inaccessibilité de la somme de données qu'il contient. Une deuxième difficulté provient de l'impossibilité d'y effectuer des consultations transversales, visant à extraire du dictionnaire un ensemble de lexèmes partageant une même propriété autre qu'étymologique (géographique, chronologique etc.). L'impossibilité d'assurer l'exhaustivité d'une recherche dans l'ouvrage est un troisième problème, auquel s'ajoute, enfin, l'inaccessibilité des mises à jour de son contenu.

Ces difficultés découragent les utilisateurs débutants et empêchent les spécialistes les plus aguerris d'exploiter l'ouvrage comme il le mérite :

> Le *FEW* se présente donc comme un ouvrage brillament conçu, un trésor d'une richesse prodigieuse, tant en quantité qu'en qualité [...] ; en même

> temps, sa structure est extrêmement touffue. Or la combinaison de ces deux caractéristiques génère une situation quelque peu paradoxale, car à l'abondance matérielle (documentation et analyses) s'oppose l'insuffisance des voies d'accès (incohérence du programme lexicographique). Inévitablement se pose donc la question de l'exploitation de ce trésor [...]. (Büchi 1996, 309)

L'informatisation du FEW pourrait apporter une réponse à ce problème. Malheureusement, les difficultés du FEW qui rendent souhaitable son informatisation font elles-mêmes obstacle à cette dernière. La masse du dictionnaire, la densité de son contenu, la spécificité de sa police de caractères et sa complexité structurelle constituent autant de particularités qui rendent le projet utopique :

> [...] le FEW constitue une somme qui permet de situer les apports nouveaux ; l'ouvrage serait encore plus utile s'il était informatisé (mais une police de caractères surabondante, en particulier pour noter les formes dialectales, en rendrait hélas le coût exorbitant). (Martin 2004, 145)

L'informatisation du FEW, pressentie comme le seul moyen de résoudre les problèmes d'accessibilité de l'ouvrage, est donc restée jusqu'à présent un rêve :

> Il est vrai que l'informatisation complète du *FEW*, telle qu'elle a pu être envisagée par certains [...], résoudrait une partie des problèmes d'accessibilité. Mais un tel projet paraît pour le moins prématuré en ce moment, même si quelques pas encore timides dans ce sens ont déjà été faits. (Büchi 1996, 309)

Notre conviction que l'informatisation du FEW constituerait une avancée percutante pour les études de linguistique historique romane a motivé la mise en chantier de cette thèse, avec pour point de départ la question suivante : est-il possible de rétroconvertir les 25 volumes imprimés en un dictionnaire informatisé ?

Disons-le d'emblée dès l'introduction de cette étude : la réponse est positive. Cette thèse prouve la faisabilité d'une informatisation du FEW capable de lever la majorité des obstacles rencontrés par ses utilisateurs. Elle propose une modélisation du discours fewien apte à rendre compte de ses particularités et à permettre son informatisation. Elle démontre que cette informatisation peut se faire de façon automatisée, malgré les incohérences structurelles de l'ouvrage.

0.1 Objectif de l'étude

En commençant cette étude, nous n'imaginions pas parvenir à un tel résultat. Notre objectif initial, qui était de préparer l'informatisation de l'ouvrage, se déclinait en trois points :

1. une étude de faisabilité, consistant à décider si, malgré les particularités structurelles et techniques qui caractérisent le FEW, il est raisonnable de s'atteler à son informatisation ;

2. une réflexion concernant les types d'interrogations (par étymons, par unités lexicales, par parlers, etc.) à prévoir, ainsi que sur les modalités d'interrogation (neutralisation des diacritiques, traitement des 188 caractères phonétiques spéciaux) ;

3. des propositions sur le traitement de l'épistructure du FEW (compléments et corrections à la fin de certains volumes) ainsi que sur celui de la parastructure de l'ouvrage (ajouts et corrections externes).

Il est rapidement apparu que les deuxième et troisième points, pour important qu'il soient, étaient dépendants du premier. Déterminer si le FEW peut être informatisé constitue le premier objectif à atteindre avant d'aller plus loin. Notre recherche s'est donc focalisée sur cette question. Or, la notion d'informatisation est assez floue : qu'entendre exactement par *informatiser* ? En vue de mieux cerner le problème, nous nous sommes demandé

1. quel devait être le but d'une informatisation du FEW ;

2. quelle forme devrait prendre un FEW informatisé pour répondre au but fixé ;

3. de quelle façon le processus d'informatisation devrait avoir lieu.

Pour répondre à la première question, il a fallu déterminer quelle utilisation était faite du FEW et en quoi l'informatisation pouvait répondre aux besoins des utilisateurs. Il a ensuite été possible de concevoir un modèle de FEW informatisé qui corresponde aux attentes de la communauté scientifique. Encore fallait-il que ce modèle soit assez réaliste pour être applicable en pratique. La densité du FEW requiert que les 25 volumes soient informatisés de façon automatisée, comme ce fut le cas pour le TLFi (cf. Dendien et Pierrel 2003). La question de la faisabilité d'une informatisation revient donc à s'interroger sur la faisabilité d'une automatisation du processus.

Cette question centrale nécessite l'élaboration d'un prototype : sans cela, l'étude risquerait de rester théorique et, dès lors, peu probante. Nous nous sommes attelée à la tâche en mettant au point des algorithmes capables d'analyser le discours fewien pour le convertir dans le modèle proposé. L'implémentation de ces algorithmes a été réalisée par un informaticien, après quoi il a été possible de les tester sur corpus et d'analyser le résultat. C'est ce résultat concret qui nous permet de répondre positivement à la question de la faisabilité. Par la même occasion, la collaboration interdisciplinaire nécessitée pour la construction du prototype a permis d'aller plus loin que les objectifs prévus : en effet, davantage qu'un prototype, cette collaboration a donné lieu, grâce à l'implémentation de tous les outils nécessaires au bon fonctionnement des algorithmes, à la mise au point d'un logiciel complet et fonctionnel de rétroconversion du FEW.

0.2 Plan de l'étude

Cette étude se décline en deux parties.

La première partie propose une modélisation du discours fewien qui soit compatible avec les objectifs de son informatisation. Elle se subdivise en trois chapitres. Le premier chapitre ($\rightarrow$ 1) fait le point sur les raisons qui justifient une informatisation du FEW et sur les tentatives qui ont été effectuées dans ce but avant que nous ne débutions

notre étude. Le second chapitre ($\rightarrow$ 2) analyse les structures du FEW en adoptant le point de vue des utilisateurs de l'ouvrage, de façon à déterminer exactement les objectifs à atteindre. Sur la base de cette analyse, le troisième chapitre ($\rightarrow$ 3) construit une modélisation du discours fewien, formalisée en XML, censée le rendre exploitable selon les modalités définies dans le chapitre précédent.

La seconde partie de notre étude examine la façon dont le modèle proposé peut être inséré de façon automatisée dans le texte fewien. Cette seconde partie se subdivise comme la première en trois chapitres. Le premier chapitre ($\rightarrow$ 4) présente l'architecture générale du logiciel (dit de rétroconversion) qui a été conçu dans ce but. Le deuxième chapitre ($\rightarrow$ 5) se focalise sur le noyau du logiciel. Nous y expliquons de quelle façon chacun des types d'information du FEW est reconnu de façon automatisée, malgré les incohérences structurelles et typographiques de l'ouvrage. Enfin, le troisième chapitre ($\rightarrow$ 6) analyse le résultat fourni par l'application du logiciel sur une centaine d'articles du FEW et examine si le balisage XML inséré de façon automatisée permettra l'exploitation des articles selon les objectifs définis dans la première partie de l'étude.

Première partie

Modélisation du FEW

Chapitre 1

État de la question

L'informatisation du FEW est un rêve évoqué, depuis quelques années déjà, par plusieurs linguistes éminents, au détour d'études portant sur des questions diverses. L'idée faisant son chemin, quelques réalisations en rapport avec l'informatisation du FEW ont été entamées à l'ATILF au début des années 2000. Ces réalisations n'ont pas toutes été couronnées de succès, essentiellement à cause de la grande variabilité du discours fewien, qui résiste à des approches trop rigides. Toutefois, elles ont eu le mérite de mettre en lumière les particularités fewiennes qui posaient problème dans le cadre d'une informatisation. Ces divers projets constituent, dès lors, un point de départ pour notre étude.

La première des réalisations atilfiennes est la publication en 2003 d'un index sélectif des formes du FEW, fruit d'un projet qui s'est étendu sur trois ans (cf. ATILF 2003). La deuxième consiste en une réflexion d'ensemble menée par une étudiante, Virginie Beckert, lors d'un stage de DESS à l'ATILF, réflexion qui a abouti à une maquette de « mini-FEW informatisé » (cf. Beckert 2003). La troisième concerne l'élaboration par Anne-Christelle Matthey (rédactrice au FEW) et Gilles Souvay (ingénieur informaticien) d'une DTD (*Document Type Definition* : une sorte de grammaire informatique, → 1.4) ayant pour but le balisage des articles de la refonte du FEW (nouveaux articles de la tranche alphabétique B-).[1] La rédaction des articles de la refonte a en outre inspiré un certain nombre d'autres propositions informatiques, destinées à faciliter l'accès au FEW en attendant son informatisation complète.

Ce chapitre propose un relevé des témoignages justifiant une informatisation du FEW. Il synthétise ensuite le contenu des divers projets qui ont été entrepris et reprend pour chacun d'eux les difficultés rencontrées.

[1]Les développements qui suivent s'appuient essentiellement sur des entretiens réalisés à l'ATILF avec les personnes qui ont mené les projets en question (voir dans la bibliographie, Interviews).

1.1 L'intérêt d'un FEW informatisé

1.1.1 L'accessibilité des données

La première difficulté émise par les utilisateurs du FEW concerne le peu de lisibilité du texte lexicographique. Les vingt-cinq volumes du FEW ne bénéficient pas, en effet, d'une mise en page aérée :

> Pour la tranche de nomenclature considérée, les 14 colonnes du tome 1er sont devenues (dans la refonte) 96 pages (et d'une densité typographique accrue, un peu « limite » pour les yeux de certains). (Rézeau 1989, 165)

L'agencement des nombreuses informations obéit à un grand principe d'économie. Les sources sont citées avec précision, mais sous une forme abrégée. De même, une information n'est pas répétée inutilement, selon la règle voulant qu'une information reste valable tant qu'elle n'a pas été remplacée.

La consultation du FEW nécessite donc une initiation préalable et exige de la part du lecteur une grande concentration. La difficulté n'est pas seulement typographique, mais surtout structurelle :

> Dans le cas du F.E.W., la difficulté de lecture est augmentée par la complexité des structures et l'abondance de l'information. [...] le caractère extensif et la qualité des références du F.E.W. rendent toujours possible son utilisation critique. Il est vrai que ces qualités ont pour pendant l'éducation nécessaire des utilisateurs. Pour chercher l'origine d'une forme lexicale dans le F.E.W., il faut connaître la nature « romane », « germanique » ou « empruntée » de son étymon et avoir des notions de phonétique historique [...]. L'étymon trouvé, il faut repérer dans un ou plusieurs sous-ensembles la forme recherchée, ce qui n'est pas toujours commode dans les grands articles [...]. (Rey 1971, 103-104)

Seule la pratique permet en fin de compte de se familiariser avec le FEW et d'acquérir le minimum de « culture fewienne » qui procure une certaine facilité de lecture. Lorsque Gilles Roques écrit que « les étudiants et le public cultivé ne peuvent pas avoir un accès très commode au FEW ou au TLF » (Roques 1991, 94), il ne désigne pas seulement l'accès au dictionnaire lui-même (qui, existant uniquement sur format papier, est consultable presque exclusivement dans des bibliothèques spécialisées), mais aussi la difficulté de se repérer à travers les structures complexes du FEW et de trouver rapidement et sûrement l'information recherchée.

1.1.2 Les consultations transversales

La visée diachronique de l'œuvre présente l'inconvénient de permettre difficilement une recherche en synchronie :

> Pour ce qui est du *FEW*, sa méthode – traitement de formes galloromanes à partir d'étymons (t. 1-20, 24-25) ou regroupement panchronique de formes à partir de concepts (t. 21-23), [...] résulte en une dispersion des unités d'un état de langue donné [...]. (Wooldridge 1998, 224)

Cette dispersion des unités analysées pose un problème de consultation lorsque l'on traite de questions particulières, relevant de grands ensembles linguistiques, tels que les déonomastiques (unités lexicales dont l'étymon est un nom propre) :

> Quand on parle de déonomastiques dans le *FEW*, il se pose le problème de l'accessibilité des données. Les matériaux déonomastiques du *FEW* ne bénéficient pas d'un traitement dans une section à part. De plus, le corpus de déonomastiques galloromans qu'on peut délimiter à l'intérieur du *FEW* n'est pas indexé (contrairement à celui du *REW*). Avec le *REW* et le *FEW*, on se trouve ainsi dans une situation inconfortable : d'un côté une matière déonomastique facilement accessible parce qu'indexée, mais pauvre [...], de l'autre un riche recueil de déonomastiques non indexé. Celui qui s'intéresse au traitement des déonomastiques dans le *FEW* doit donc d'abord en constituer la nomenclature, qui a la propriété d'être cachée. (Büchi 1992, 69-70)[2]

La question des déonomastiques est représentative d'un type particulier de consultation du FEW : la lecture transversale. L'objectif est de mettre en évidence, à travers tous les articles du dictionnaire, un ensemble linguistique spécifique tels que les régionalismes[3], le vocabulaire français emprunté aux dialectes d'oïl[4] ou encore le lexique français du XVIe siècle[5]. L'organisation des données dans le FEW ne permet que malaisément ce type de recherche :

> Dans le *FEW* [...], les unités du lexique français se trouvent dispersées, d'abord par la nomenclature, fondée sur les étymons, ensuite par la microstructure, qui procède par regroupement de formes dialectales dont le français ; dans les volumes onomasiologiques, la rencontre d'unités du lexique français d'une synchronie donnée n'est que fortuite. Dans ces conditions, le FEW peut difficilement rendre compte de familles lexicales fonctionnelles ou de microsystèmes morphologiques ou sémantiques. (Wooldridge 1998, 211)

De même, alors que le FEW possède la grande qualité de signaler ses sources et donne ainsi au lecteur les moyens d'une utilisation entièrement critique, il est impossible de connaître de façon exacte l'importance absolue et relative des sources utilisées :

> La typologie, donnée ci-dessus, des attestations *EN* du *FEW* montre – du moins en ce qui concerne le tome 10 – que Thes n'est pas utilisé par le *FEW*. Note de bas de page : K. Baldinger, ancien collaborateur du FEW, **pense** que *Thes* n'aurait été cité en fait qu'une seule fois dans tout le FEW

[2]Cf. aussi Büchi 1991, 139 : « Le *FEW* constitue à ce jour le principal recueil de déonomastiques galloromans. Mais comme il n'est pas indexé [...], ce recueil n'est accessible qu'à travers une investigation préalable. » La publication récente d'un index sélectif des formes répertoriées par le FEW (ATILF 2003) a apporté une première réponse à ce problème.

[3]Cf. Lagueunière 1998.

[4]Cf., pour une étude de cette section du lexique français et son traitement dans le FEW, Gebhardt 1982 (à nuancer par Chambon 1997).

[5]Le reflet de cet ensemble lexical dans le FEW et dans les versions du *Grand Dictionnaire françois-latin* forme l'objet de Wooldridge 1998.

> (Baldinger K., « Estienne 1531 et son importance pour l'histoire du vocabulaire français », *Wolfenbütteler Forschungen*, 18 (1982), 10-11).[6]

En règle générale, établir des comptages ou des statistiques dans le FEW (en répertoriant, par exemple, le nombre de formes se rapportant à un étymon ou le nombre d'unités lexicales présentant un suffixe donné) est une entreprise impossible.

Parce que la dispersion des données rend leur accès difficile, toutes ces interrogations transversales, notamment celles qui portent sur de grands ensembles linguistiques, nécessitent actuellement une lecture approfondie et fastidieuse de nombreuses pages du FEW.

1.1.3 L'exhaustivité des résultats

La présence dans le FEW de formes « cachées » est une réalité dont il faut constamment tenir compte lorsqu'on le consulte. À l'instar de celui d'Éva Buchi sur les déonomastiques (Büchi 1992), nombre de travaux se heurtent à ce problème. Il est impossible d'être certain qu'une forme ne se trouve pas dans le dictionnaire. Kurt Baldinger recommande d'ailleurs, non sans humour, de « lire le FEW d'un bout à l'autre » avant de prétendre qu'un mot y manque ![7] Par conséquent, affirmer qu'un mot fait défaut dans le FEW nécessite quelques précautions discursives :

> Dans les documents, nous avons pu relever 24 mots qui manquent dans le *FEW*. Puisqu'il s'agit exclusivement de lexèmes dont les étymons sont parfaitement connus, et que, par conséquent, les articles concernés sont aisément repérables, nous ne pensons pas qu'une attestation cachée ait pu nous échapper. Toutefois, pour permettre des vérifications, nous indiquons [...] l'endroit du *FEW* où le mot devrait figurer. (Schmitt 1992, 310)

Cette règle de prudence, qui consiste à toujours signaler l'endroit du FEW où l'on a cherché, est généralement respectée :

> En afr. par contre, *determinaison* « détermination » est bien attesté (Gdf, II, 686c ; T.L. II, 1831) : mais le *FEW* (III, 57a, *determinare*) ne le donne pas. [...]
>
> *Fendoison* 'fente'. Gr. d'Hauterive est le seul à citer ce mot. Il n'est pas dans le *FEW* (III, 549-551, *findere*). (Merk 1980, 283 ; 287)

> GDFL Rudenter, *en terme de menuiserie, & architecture, ainsi l'on dict rudenter, une colomne, & une colomne rudentée.* (M, dB s.v. RUDENTER)
>
> *Rudenté, colomnes rudentees.* (P s.v. RUDENTE)
>
> FEW Ø ? (étymon *rudens* Ø) (Wooldridge 1998, 254)

La difficulté d'accéder à l'exhaustivité d'une recherche dans le FEW ne concerne pas seulement les formes-objet, que l'index publié en 2003 permet maintenant de repérer plus facilement, mais aussi les étymons cachés. Un grand nombre d'étymons

[6] Wooldridge 1990, 252 (c'est nous qui mettons en gras).

[7] Baldinger 1974, 25.

(notamment des dérivés) n'ont en effet pas droit de cité dans la macrostructure, mais se trouvent relégués en tant que sous-lemmes dans des articles consacrés à d'autres étymons (cf. Büchi 1996, 52-73). Le relevé effectué par Éva Buchi (Büchi 1996, 405-564) ne compte pas moins de 5800 étymons cachés ! En outre, une centaine d'entre eux apparaissent dans plusieurs articles, où ils bénéficient d'un traitement inégal (Büchi 1996, 562-4).

Il s'y ajoute un autre problème, celui des nombreuses corrections effectuées par von Wartburg lui-même ou ses collaborateurs : « Quelquefois ces corrections se trouvent cachées quelque part dans la forêt vierge du *FEW*, où elles sont pratiquement introuvables [...]. » (Baldinger 1993, 514)

Le FEW possède en effet cette grande qualité de se corriger lui-même, au fur et à mesure de sa rédaction. Le commentaire situé à la fin des articles est un lieu privilégié de discussion, qui peut remettre en cause des analyses publiées antérieurement :

> Étant donné sa longue durée de publication, le FEW a la possibilité, sinon l'obligation, d'introduire des auto-corrections dans ses colonnes ; celles-ci sont introduites soit dans le commentaire des articles, soit dans les notes [...]. Ces corrections internes concernent essentiellement de faux classements antérieurs [...]. (Büchi 1996, 156-157)

À la suite de cette remarque, Éva Büchi relève dans le FEW, parmi d'autres, l'exemple d'auto-correction suivant :

> Als erbwort ist das hier 1,10 behandelte afr. *esterdre* unter EXTĔRGĔRE einzureihen, hier 3, 327, wo nachzutragen sind apr. *esters* „pur, sans mélange", Puiss. „excepté". (von Wartburg in FEW 24, 57a, ABSTERGERE)

Un cas de figure plus radical et très fréquent est constitué par une partie d'article qui remplace une partie d'un autre article, en classant des matériaux semblables sous un étymon différent, sans toujours préciser l'existence de l'article antérieur. Il s'agit du problème des doubles classements, phénomène dont l'existence a été mise en évidence par Kurt Baldinger (1980) et dont les réalisations ont été traquées avec bonheur par un groupe de lexicologues fédérés par Jean-Pierre Chambon[8]. Ces classements multiples (souvent doubles ou triples, voire davantage) deviennent dangereux dans la mesure où ils ne proposent pas les mêmes matériaux ou les mêmes rattachements étymologiques, car le lecteur qui ne consulte qu'un seul des articles risque de passer à côté d'informations importantes :

> Il arrive aussi, mais c'est exceptionnel, qu'une même famille ait été traitée deux fois, avec les mêmes matériaux ou avec des matériaux complémentaires, sous deux étymons distincts. Le plus souvent, la solution étymologique reste fondamentalement la même [...]. Bien que l'inconvénient soit plus grave pour le consulteur (qui, bien souvent, à la recherche d'une forme ou d'une famille, ne lit pas pour autant tout l'article du *FEW*, malgré le sage conseil d'Alain Rey, et encore tout le dictionnaire !), on reste alors dans le cadre des erreurs matérielles, de nature simplement dictionnairique. Mais il se trouve aussi que les étymologies proposées soient différentes. (Chambon *et al.* 1987, 168-169)

[8]Cf. Chambon *et al.* 1987 ; Brochard *et al.* 1989 ; Boutier *et al.* 1990, 1992, 1994 ; Chambon *et al.* 1999.

Un important travail de traque a donc commencé afin de repérer toutes ces « étymologies doubles » :

> 288. BOUCACOUI (article à biffer 1, 471b) = MOKAEM (20, 72b-73a). – L'article du volume 20 remplace visiblement celui du volume 1, mais rien n'en avertit le lecteur. [M.-J. Brochard *in* Boutier *et al.* 1992, 394].
>
> 345. JUK- (5,62b) = JUCKEN (16, 287b). – [...] les matériaux sont presque complètement différents (sauf une forme de GrCombe) et il n'y a aucun renvoi explicite d'un article à l'autre. [Chambon *in* Boutier *et al.* 1992, 403].
>
> 413. [...] Ce groupe est absent de la première version de l'article VES- (14, 339b) ; c'est à cette première version que renvoie von Wartburg 14, 563b et n. 11, sous VITIUM. [...] [Chambon *in* Boutier *et al.* 1992, 414].

L'absence de renvois réciproques, sans cesse signalée par ces travaux, montre combien la consultation du FEW est périlleuse. Malgré les nombreuses corrections apportées à ce jour, le manque de certitudes quant à l'exhaustivité d'un résultat de recherche dans le FEW reste un problème majeur pour le chercheur en lexicologie historique.

1.1.4 La mise à jour du dictionnaire

Les corrections au FEW ne concernent pas seulement l'organisation interne du FEW, mais aussi l'ajout d'informations nouvelles, qui remplacent ou corrigent celles du dictionnaire :

> En fait, toute liste de datations nouvelles, toute hypothèse étymologique, toute enquête dialectologique constitue une addition au *F.E.W.*, et une mise au point partielle de ce grand dictionnaire. (Rey 1971, 83)

Ce constant travail de remaniement prend, pour toutes les raisons citées plus haut, un temps considérable. Von Wartburg était déjà confronté à ce problème :

> Une autre difficulté, ce sont les nouveaux matériaux : c'est un problème très délicat et difficile que celui des dictionnaires de patois qui paraissent et des nombreuses éditions de textes. [...] Comment voulez-vous qu'avec le peu d'années qui me restent, je puisse incorporer tout cela aux matériaux existants du *FEW* ? C'est impossible.[9]

L'impossibilité purement pratique de rééditer le dictionnaire en y incorporant les données nouvelles et les remises en question a donné naissance à une littérature annexe au FEW (sa « péristructure »), qui recense un grand nombre de mises à jour. Le FEW se retrouve ainsi au centre d'une importante production scientifique englobant tant les comptes rendus des différents fascicules, des études de lexicologie historique proposant des ajouts et corrections à l'ouvrage (la mention «à ajouter au FEW tome X page Y» étant devenue un classique du genre) que des monographies spécialement dédiées à

[9] von Wartburg 1961, 212 (cité par Pfister 1980, 133).

l'amélioration de certaines parties de l'œuvre[10]. Un bel exemple de ce « genre textuel » est Arveiller 1999 (publication posthume reprenant une série de 25 articles parus entre 1969 et 1996), où sont rediscutées les étymologies du tome 19, dédié aux *Orientalia*, à la lumière d'un important corpus d'attestations nouvelles[11].

Les corrections effectuées peuvent parfois s'avérer radicales, d'où leur importance pour l'utilisateur du dictionnaire – qui, malheureusement, ne possède aucun avertissement signalant que l'article qu'il lit dans le FEW a été revu et corrigé, voire même supprimé : « Der Artikel *LAIDA „Weg" (FEW und REW) ist zu streichen » (Baldinger 1968, 56)

1.1.5 Appels à l'informatisation du FEW

En voyant le nombre de publications récentes apportant explicitement une mise à jour au FEW, on se prend à espérer une solution facile :

> [...] Face à ce grouillement, on peut toujours rêver à un dictionnaire total à la Jorge-Luis Borges (« La bibliothèque de Babel »), inlassablement augmenté, remanié, élagué et reclassé par des machines infaillibles, et que tout un chacun pourrait non seulement consulter, mais enrichir de ses trouvailles en tripotant, de chez lui, quelques boutons. (Gebhardt 1982, 33)

C'est ainsi que la communauté internationale des chercheurs en lexicologie historique a peu à peu osé le rêve d'une informatisation du FEW :

> La rédaction du *FEW* s'achève ; la tâche immense mais logique qui s'impose dorénavant est de l'informatiser, en y intégrant, sans briser la méthode mais en permettant d'autres lectures des données, toutes les corrections et compléments sûrs qui ont été apportés depuis plusieurs décennies et en y ajoutant ceux à venir. (Wooldridge 1990, 239)

En 2004, Marcello Aprile émet les mêmes désirs que Wooldridge, lorsqu'il compare la situation du LEI à celle du FEW :

> I tempi della costruzione di un archivio del genere appaiono invece già pienamente maturi per il FEW (anche se la discussione su questo tema si è appena aperta), per il quale la pur faticosa informatizzazione renderebbe possibile insieme più obiettivi : l'eliminazione del fastidioso problema delle etimologie doppie o triple (Baldinger 1993), l'aggiornamento, la revisione e la razionalizzazione del materiale esistente ora parzialmente disperso tra i volumi dell'opera wartburgiana. (Aprile 2004, 227)

Cette attente d'une informatisation ne concerne pas seulement les spécialistes de l'étymologie galloromane. Le FEW constitue une référence unique et précieuse pour tous ceux qui s'intéressent non seulement à l'histoire des quatre langues galloromanes, mais aussi à celle des autres langues romanes :

[10] La série d'articles consacrée au dépistage des « étymologies doubles » (→ 1.1.3) fait également partie de la péristructure du FEW.

[11] Voir aussi Baldinger 1988, 1998, 2003.

> Par ses dépouillements inédits, le *FEW* a un intérêt incontestable pour les recherches romanes en dehors du domaine galloroman[156] ; dans beaucoup de cas, ses articles présentaient à leur époque (et présentent parfois encore aujourd'hui) le panorama le plus étendu (par exemple pour l'italoroman avant la parution de l'article correspondant du *LEI*). Bien souvent, il arrive que le *FEW* résolve des problèmes étymologiques d'autres domaines romans[157]. (Büchi 1996, 142, qui cite Zamboni 1976)

Malheureusement, les difficultés de lecture évoquées ci-dessus font que le FEW est peu exploité, voire peu connu, par les linguistes et les philologues, notamment en France, où l'on a tendance à ignorer cet ouvrage majoritairement rédigé en allemand [12]. Dans son relevé des conditions déterminant le renouveau des études en linguistique historique française, Robert Martin rend pourtant hommage au FEW dans des termes non ambigus :

> *Conditions pour un nouvel essor de l'histoire interne du français.* – On notera tout d'abord l'accroissement considérable des sources (et leur progressive informatisation) ; [...] des ouvrages de synthèse recueillent et organisent une masse prodigieuse d'informations ; c'est le cas tout particulièrement du dictionnaire étymologique (*Französisches Etymologisches Wörterbuch*) de W. v. Wartburg [...] ; le FEW constitue une somme qui permet de situer les apports nouveaux ; l'ouvrage serait encore plus utile s'il était informatisé (mais une police de caractères surabondante, en particulier pour noter les formes dialectales, en rendrait hélas le coût exorbitant). (Martin 2004, 145)

Les inquiétudes émises ici par Robert Martin concernant l'impossibilité d'informatiser le FEW sont révélatrices. Ce projet apparaît de prime abord comme une utopie, une entreprise gigantesque, à la hauteur des ambitions d'un von Wartburg. Son utilité est néanmoins certaine. Outre l'avantage d'apporter un gain de temps inestimable dans les recherches, un FEW disponible sous forme électronique résoudrait sans doute les différents problèmes cités ci-dessus. Il pourrait en effet

1. élargir son accessibilité, en se mettant à la disposition de tous à la fois d'un point de vue pratique (accès internet) et d'un point de vue scientifique (lisibilité du contenu) ;

2. permettre des consultations transversales, grâce à de multiples possibilités de requête :

   > La seule façon de mettre au jour ce qui concerne le français du XVIe siècle dans le *FEW* serait d'informatiser les 25 volumes... puis de les interroger à partir de repères comme « fr. », « mfr. », « 16e s. », etc. (Wooldridge 1998, 211)

3. vérifier l'exhaustivité des résultats obtenus, en permettant l'extraction d'informations qui restaient cachées dans la version papier ;

[12]Cf. Plouzeau 2000, 509-510, qui se plaint de la sous-exploitation du FEW dans les universités françaises, et Büchi 1996, 137 : « La métalangue des commentaires rédigés par Wartburg et par ses collaborateurs germanophones est en principe l'allemand. On connaît les grands problèmes de compréhension et, partant, de sous-exploitation que ce choix a entraînés pour l'ouvrage [...] ».

4. et, enfin, faciliter les mises à jour qui s'imposent :

> Les avantages de la *rétroconversion* d'un dictionnaire imprimé en *dictionnaire informatisé* sont considérables. En effet, la transformation du contenu d'un dictionnaire en base de données relationnelle permet d'accéder à chaque élément constitutif du texte, la structure logique de celui-ci ayant été rendue explicite. Non seulement le dictionnaire informatisé représente au sens propre le « *multidictionnaire* » par les innombrables types de lectures-consultations qu'il rend possibles, mais c'est une *base éditoriale* qui facilite les corrections, les mises à jour, les extractions, etc. [...] (Quemada 1991, 19-20)

1.2 L'index sélectif des formes du FEW

1.2.1 L'intérêt d'un index du FEW

Le but d'un index est d'aider le lecteur à trouver plus rapidement, dans un document donné, l'endroit où se trouve l'information qu'il cherche. Dans un dictionnaire habituel, où les unités linguistiques sont présentées par ordre alphabétique, il n'est nul besoin d'un index. Ce n'est pas le cas du FEW : l'organisation macrostructurelle du dictionnaire, conformément à la visée génétique de von Wartburg, regroupe les matériaux d'une même famille étymologique dans un article dont le titre correspond à l'étymon (éventuellement lointain) commun. Selon ce classement, la lecture d'un article part de l'étymon et permet de prendre connaissance de toutes les unités qui s'y rattachent. Or, les lecteurs, ou plutôt les consultants, du FEW sont généralement confrontés au cas inverse : ils sont en présence d'une unité lexicale dont ils recherchent l'étymon. Les formes apparaissant dans les articles à la suite les unes des autres, dans un ordre géo-historique et non alphabétique, il est impossible dans ce cas de consulter le FEW comme un dictionnaire habituel.

L'indexation pallie donc un handicap majeur du FEW, en permettant au lecteur de trouver, dans les 25 volumes du dictionnaire, l'endroit où a été classée l'unité qui l'occupe[13]. Bien conscient de cette nécessité, von Wartburg avait le projet, malheureusement jamais abouti, de réaliser un index complet de son œuvre :

> Comme on le sait, chaque tome du FEW se termine par un index des mots traités. Il n'a jamais été envisagé de cumuler ces différents index, au contraire leur caractère partiel et provisoire était régulièrement réaffirmé, en même temps que le projet d'un index complet qui serait mis en chantier dès la fin de l'ouvrage. Quand Wartburg s'engagea, au cours des années soixante, dans la refonte des premiers tomes, du même coup il repoussait cet index général aux calendes grecques : comment indexer un ouvrage en plein remaniement ? La nécessité de cet index général ne s'en faisait pas moins sentir. (Chauveau 2006, 34)

[13]Comme le précise ATILF 2003, VII, cette aide ne remplace évidemment pas la lecture complète de l'article en question : « [...] une bonne utilisation du FEW comporte une lecture (au moins cursive) de l'article complet dont relève la lexie à laquelle on s'intéresse [...]. Nous invitons donc les lecteurs à ne jamais se contenter de localiser la lexie recherchée dans une page précise, sous peine de passer à côté d'informations primordiales ».

En 1998, l'INaLF a décidé de combler ce manque. L'index, publié en 2003, a reçu la contribution d'une trentaine de spécialistes, émargeant notamment à l'ATILF (laboratoire issu entre-temps de la fusion de l'INaLF et du LanDisCo), mais aussi aux universités de Heidelberg (DEAF), Liège, Neuchâtel, Strasbourg et Zurich[14].

1.2.2 Le choix des informations à indexer

Il s'agit d'un index sélectif : sur les quatre ou cinq millions d'unités lexicales répertoriées par le FEW, l'index ne reprend que 275 295 formes, considérées comme les plus représentatives. Cet ensemble restreint remplit malgré tout deux tomes et 2 370 pages. Les critères qui ont permis de juger de la représentativité d'un lexème sont détaillés dans l'introduction à l'index (ATILF 2003, IX-X) et résumés par Jean-Paul Chauveau comme suit :

> Le français étant la « langue-toit » du domaine galloroman, les formes du français moderne ont la priorité. Celles qui appartiennent à un état ancien du français ou à une autre langue ou un dialecte galloromans ne sont enregistrées que si elles diffèrent de la forme française par leur première syllabe ou par l'initiale de leur deuxième syllabe ou bien si le type n'existe pas en français moderne. On n'a pas non plus retenu les dérivés suffixaux dont la forme interne est transparente à tout francophone contemporain. (Chauveau 2006, 34)

La sélection a donc été guidée par un principe pratique, consistant à permettre au lecteur de retrouver n'importe quelle unité galloromane citée par le FEW à partir de celles de l'index, en prenant en compte la contrainte d'une consultation par ordre alphabétique. Il faut préciser qu'ont été exclues d'office les unités autres que galloromanes, y compris celles en latin médiéval, ainsi que les unités galloromanes citées dans les commentaires uniquement ; en outre, les articles des premiers volumes qui ont bénéficié d'une refonte n'ont évidemment pas été indexés (ATILF 2003, V).

Outre les formes simples du français moderne et les formes anciennes ou dialectales qui en diffèrent par le début du mot[15], l'index retient donc tous les préfixés, puisqu'ils diffèrent eux aussi des formes simples par leurs premières lettres. Les suffixés sont quant à eux retenus uniquement lorsque leur dérivation n'est pas triviale en français moderne, c'est-à-dire lorsque leur base est difficilement reconnaissable (par exemple *heureux* < afr. *heur*), lorsque le suffixe n'est pas productif en français contemporain, ou encore lorsque leur formation est phonétiquement irrégulière (toujours du point de vue du français contemporain)[16]. Enfin, les composés dont les éléments sont séparés par un blanc ou un trait d'union sont tous indexés, et classés sous tous leurs éléments lexicaux[17].

[14]Les noms des collaborateurs sont cités dans ATILF 2003, III.

[15]Sont donc exclues dans tous les cas les variations phonétiques de la fin du mot, ainsi que les variations graphiques récurrentes du français.

[16]Ces précisions, ainsi que les suivantes, proviennent d'un document fourni par Éva Buchi aux collaborateurs en septembre 1998 et énonçant les règles de sélection à appliquer. L'auteure illustre le cas des suffixés dont la dérivation est phonétiquement irrégulière par l'exemple de ang. *rubiette* f. « rouge-gorge » < fr. *rubi* m. « rubis » : si la dérivation s'était faite aujourd'hui, on s'attendrait à **rubisette*.

[17]Les locutions n'apparaissent pas dans l'index publié. Elles ont cependant été traitées en vue d'une publication sur Internet ou en cédérom (non aboutie en 2006), selon les principes suivants : les locutions ont été

Pour chaque forme indexée, trois informations sont relevées : l'étymon sous lequel elle est classée, sa localisation précise dans le dictionnaire (tome, volume éventuel, page et colonne) et son appartenance à une variété géo-historique du galloroman, sous la forme d'un sigle géo-historique qui n'est pas obligatoirement celui que donne le FEW, mais constitue dans bon nombre de cas une localisation approximative (cf. ATILF 2003, VI et ci-dessous I.4). L'index présente les informations comme suit :

abais (poit.) abbatuĕre 24, 18a
abáiso (cév.) wahsjan 17, 451a
abaissier (afr.) *bassiare 1, 273a

1.2.3 Les étapes de la réalisation

1.2.3.1 La constitution d'un fichier numérisé et balisé

Les trois premières étapes de la réalisation furent les suivantes :

1. Sélection des formes à indexer : le FEW fut scindé en plusieurs lots, confiés aux 27 collaborateurs qui lisaient et surlignaient sur papier, selon les critères déterminés, les unités lexicales à retenir ainsi que les étiquettes chronologiques ou géolinguistiques correspondantes[18].
2. Numérisation : pendant ce temps, le FEW fut numérisé à l'ATILF par Hassen Hadj Ammar (assistant ingénieur au CNRS). Le résultat de l'océrisation était très imparfait, mais suffisant pour permettre l'étape suivante.
3. Balisage manuel : le balisage (sur le fichier numérisé) de chaque forme et des informations connexes a été réalisé par huit personnes, au moyen de programmes écrits par F. Henry (ingénieur de recherche au CNRS), qui permettaient de sélectionner le contenu autour duquel le logiciel insérait lui-même les balises ouvrantes et fermantes.

Les balises étaient les suivantes :

– <ved> </ved> pour l'étymon-titre (vedette),
– <geo> </geo> pour le sigle géolinguistique,
– <frm> </frm> pour l'unité galloromane (forme),
– <pag> </pag> pour le numéro de page,
– <col> </col> pour la lettre (<a> ou <b>) de la colonne[19].

exclues lorsque l'élément étymologisé constituait l'initiale de la locution ou en était l'élément sémantique central ; des locutions appartenant à ce dernier cas ont cependant été indexées lorsqu'elles contenaient un autre élément jugé intéressant (par exemple, s.v. ARMA, *faire ses* premières *armes* ou encore passer *l'arme à* gauche).

[18] Une seule étiquette par forme. En cas d'étiquette multiple, ils avaient pour consigne de surligner la plus générale (« occit. » plutôt que « lang. », « lang. » plutôt que « Alès ») ou, dans le doute, celle qui précédait immédiatement la forme.

[19] Le contenu des deux dernières balises était à insérer manuellement dans le texte à l'endroit qui correspondait à chaque début de page et de colonne dans le FEW.

Si plusieurs lexèmes correspondaient à une indication géohistorique commune, un numéro était attribué aux balises <geo> et <frm>, comme suit :

- <geo 1-n> </geo 1-n>
- <frm 1> </frm 1>
- <frm 2> </frm 2>
- ...
- <frm n> </frm n>

L'orthographe des objets balisés (parfois accidentellement modifiée lors de la numérisation) était par la même occasion vérifiée et corrigée. Ces trois premières étapes ont donné naissance à un fichier balisé, sur lequel Françoise Henry est intervenue afin de confectionner l'index proprement dit.

1.2.3.2 La constitution de l'index

> Une fois le texte du FEW numérisé, les formes retenues et leurs éléments connexes nécessaires (indicateur géo-historique, vedette et localisation), balisés manuellement, ont été extraits et regroupés en lignes d'index au moyen de petits programmes en *lex* et en *C++*. Les opérations de substitution d'indicateurs et surtout de tri (l'alphabet spécifique du FEW étant beaucoup plus complexe que l'alphabet standard du français) ont été faites en utilisant les fonctions correspondantes du logiciel TUSTEP, mis au point à l'Université de Tübingen. (ATILF 2003, III n.1)

Les étapes suivies pour l'indexation furent les suivantes :

Correction du balisage effectué

Chaque lot fut enregistré dans un fichier au format texte DOS, le nom du fichier indiquant le tome du FEW et le numéro du lot dans la succession de ceux découpés dans le tome. Dans chaque fichier furent appliqués successivement trois programmes (écrits en lex) de détection des erreurs de balisage. Le premier vérifiait la bonne succession et la correspondance deux par deux des balises ouvrantes et fermantes ; le deuxième, la succession correcte des balises <geo> et <frm> ; le dernier vérifiait que chaque balise possédait un contenu. Les corrections nécessaires étaient effectuées après l'application de chacun de ces trois programmes.

Extraction des objets balisés

Après l'ajout d'une balise <vol> indiquant le numéro de tome et de volume éventuel de chaque fichier, les objets balisés furent extraits (avec leurs balises) au moyen d'un programme en lex, dans l'ordre suivant : volume, page, colonne, sigle géolinguistique, forme(s), vedette. Lorsqu'un sigle géolinguistique correspondait à plusieurs unités lexicales, il fallait l'attribuer à chaque forme séparément : cette opération de réduplication des indicateurs a été effectuée au moyen d'un programme en C++.

L'extraction des objets s'est accompagnée de diverses corrections (au moyen de programmes en lex) : suppression des signes de fin de paragraphe et des coupures de mot apparaissant à l'intérieur des formes balisées, suppression des blancs en début de forme ou d'indicateur ainsi que des signes de ponctuation apparaissant en fin de forme, substitution des codes $042, $062 et $084 par les caractères correspondant (<þ>, <ñ> et <õ>), détection des confusions entre les caractères l (lettre) et 1 (chiffre).

Fabrication des lignes d'index

La fabrication des lignes d'index a été réalisée à l'aide d'un programme en C++, afin que les informations se succèdent dans l'ordre suivant : forme, indicateur, vedette, localisation (volume, page, colonne). Une série de corrections fut ici aussi nécessaire : outre la suppression des blancs en fin de lexème, d'indicateur et de vedette, une distinction fut opérée dans le balisage des vedettes entre les étymons et les concepts (pour les matériaux d'origine inconnue ou incertaine, tomes 21-23) et des macros furent exécutées afin notamment d'insérer les caractères phonétiques spéciaux du FEW, préalablement créés par Françoise Henry et enregistrés sous le nom de police « FEWr ».

Après création de l'index, le résultat fut envoyé à la relecture, et les corrections et modifications y furent intégrées.

Extraction et modification des indicateurs géo-historiques

Les auteurs de l'index ont décidé de ne pas conserver dans tous les cas le sigle géolinguistique fourni par le FEW. En effet, les lexèmes retenus dans l'index sont souvent représentatifs de tout un ensemble d'unités qui partagent des propriétés formelles (surtout phonétiques), c'est-à-dire qu'ils concernent un ensemble de formes particulières recensées par le FEW ; il était dès lors important que l'indicateur géo-historique fourni par l'index reflète le mieux possible cette représentativité et qu'il guide le lecteur vers la forme la plus proche de celle recherchée, notamment dans les cas d'homonymie ou de paronymie (cf. ATILF 2003, VI). Il a donc été décidé de normaliser les sigles, en ramenant les localisations d'extension étroite à des sigles d'étendue plus vaste. Cette opération a été effectuée en quatre étapes : (1) extraction et tri des indicateurs contenus dans l'index ; (2) constitution de la liste de ces indicateurs ; (3) définition du substitut éventuel à affecter à chaque indicateur de l'index (réalisée par Éva Buchi et Carole Champy ; pour les choix retenus, cf. ci-dessous, → 1.2.4) ; (4) substitution des indicateurs, au moyen d'un programme TUSTEP.

1.2.4 Les problèmes posés par l'indexation du FEW

L'indexation du FEW a mis en lumière un certain nombre de difficultés quant au traitement informatique de ce dictionnaire. La numérisation a révélé un premier problème de taille : le traitement des 188 caractères phonétiques spéciaux du FEW, qui n'existaient dans aucune police sur PC[20]. Le problème a été provisoirement résolu le problème en créant les caractères rencontrés dans les formes à indexer et en constituant de la sorte une police nommée « FEWr ».

[20]Une police de caractères conçue tout spécialement pour le FEW avait été créée en 1995 sous le nom *TimesIreg*, mais était utilisable uniquement sous MacOS.

a	á à â ā ā́ ă ắ ä ä́ ǟ ǟ́ å ǻ å̄ å̄́ ą́ α
ã	ã́ ã̈ ã̆ ã̆́ ã̄ α̃
æ	æ̂
b	β ƀ
c	ç
č	
d	ḍ ḏ δ đ
e	é è ê ē ḗ ĕ ĕ́ ẹ ẹ́ ẹ̄ ẹ̄́ ẹ̆ ẹ̆́ ę ę́ ę̄ ę̄́ ę̆ ę̆́ ë ê̄ ə ə́ ə̆
ẽ	ẽ́ ẹ̃ ẹ̃́ ẽ̆ ę̃ ə̃ ə̃́
f	
g	ġ ğ γ
h	ʿh ẖ ḥ
i	í ì î ī ī́ ĭ ĭ́ ị ị́ i̯ į į́ į̄ į̄́ ï
ĩ	ĩ́
j	
k	ḳ
l	λ
ł	
m	
n	ɳ
ñ	
o	ó ò ô ō ṓ ȍ ŏ ŏ́ ŏ̀ o̯ ọ ọ́ ọ̄ ọ̄́ ọ̆ ọ̆́ ǫ ǫ́ ǭ ǭ́ ǫ̆ ǫ̆́ ö ȫ

õ	ṍ ọ̃ ǫ̃ ǫ̃́
œ	œ́ œ̂ œ̄ œ̄́ œ̆ œ̆́ œ̣ œ̣́ œ̣̄ œ̣̄́ œ̣̆ œ̣̆́ œ̨ œ̨́ œ̨̄ œ̨̄́ œ̨̆ œ̨̆́ ø ø̂
œ̃	œ̃́ œ̣̃ œ̨̃́
p	
q	
r	ṛ
s	ṣ ś s̥
š	ş
t	ṭ ṯ ţ
θ	þ
u	ú ù û ū ŭ ū̆ u̯ ŭ ü ǘ ǖ ǖ́ ü̆ ụ̈ ụ̈̄
ũ̈	ũ̈́
u	ú ū ū́ ŭ ŭ́ ů ụ ų ų́ ụ̄ ụ̄́ ụ̆ u̯
ũ	ṹ
v	
w	ẅ
x	
χ	
y	ý ȳ ÿ ŷ
ỹ	
z	ź ẓ
ž	

Les signes suivants ont été neutralisés dans le tri : ( ) ' · ʿ ʾ

FIGURE 1.1 – Ordre alphabétique des caractères phonétiques

Les caractères phonétiques intervenant également dans le tri alphabétique des entrées de l'index, une deuxième difficulté a consisté à déterminer l'ordre alphabétique des divers signes diacritiques. Une fois ces critères établis, le tri a été effectué à l'aide d'un programme TUSTEP. L'ordre des caractères retenu est repris dans la figure 1.1 ; les caractères qui se trouvent sur une même ligne sont considérés comme équivalents pour le tri (ATILF 2003, VIII).

Enfin, l'indexation a mis en lumière l'incroyable variété[21] de sigles géolinguistiques présents dans le FEW, bien connue des utilisateurs :

> Une particularité de l'enchaînement des différents parlers (...) est constituée du fait que les désignations de dialectes (comme « pic. » pour ‹picard›) sont mêlées sans distinction aucune aux désignations de sous-dialectes, qui n'embrassent souvent qu'une localité précise (comme « St-Pol » renvoyant à Saint-Pol-sur-Ternoise dans le Pas-de-Calais). Il en résulte une difficulté de lecture non négligeable, ce qui n'a pas manqué d'être relevé par la critique. (Büchi 1996, 114)

[21]La refonte (dont la métalangue est le français) comptabilise 52 anciennes langues + fr., mfr., frm. et 1 277 étiquettes géolinguistiques pour les dialectes modernes. Il faut y ajouter, pour le reste du FEW, les nombreux parallèles allemands.

La normalisation de ces sigles a été réalisée par Carole Champy (ingénieur d'étude au CNRS) et répertoriée informatiquement par Françoise Henry. Les 4 130 items initialement répertoriés ont ainsi été ramenés à 79 sigles génériques (ATILF 2003, VI-VII), correspondant à des ensembles dialectaux de moyenne importance : des localisations précises comme « Famenne » ou « Faymonv. » ont été ramenées à « wall. », « St-Pol » est devenu « pic. » et « ametz. » a par exemple été modifié en « alorr. ».

Les deux obstacles majeurs que constituent, d'une part les caractères phonétiques et leurs signes diacritiques, d'autre part la diversité des sigles dialectaux, devront à nouveau être pris en compte lors de l'informatisation de l'ensemble du FEW.

1.3 Le prototype de FEWi sur Access

Une étudiante de l'Université de Metz, Virginie Beckert, a réalisé un stage à l'ATILF dans le cadre d'un DESS en Industries de la langue (mention *Traitement automatique et Techniques de traduction*). Sa recherche, effectuée d'avril à juillet 2003 sous la direction de Jean-Paul Chauveau, portait sur « L'Informatisation du *Französisches Etymologisches Wörterbuch* » et a abouti à un mémoire d'une soixantaine de pages (Beckert 2003), consultable à l'ATILF.

Nous synthétisons ci-dessous le contenu de ce travail. Afin d'en retirer des enseignements pour notre propre recherche, nous étudions ensuite les avantages (peu nombreux) et les inconvénients (majoritaires) des solutions envisagées, réalisant ainsi le souhait exprimé par l'auteure : « L'informatisation de l'intégralité du dictionnaire n'est pas à l'ordre du jour. Cependant, la réflexion faite ici peut servir d'étude préalable même si une réalisation matérielle n'est pas prévue à moyen terme » (Beckert 2003, 8).

1.3.1 Synthèse

> Mon sujet de stage a consisté d'une part, à étudier les matériaux lexicographiques et les structures du dictionnaire, d'autre part, à imaginer les différentes recherches pouvant être effectuées sur le dictionnaire s'il était converti en dictionnaire électronique. [. . .] J'ai essayé de dégager de cette réflexion un modèle de FEW informatisé (FEWi si l'on peut dire) [...]. (Beckert 2003, 5)

Le mémoire présente successivement une analyse structurelle des articles, des propositions d'interrogations et, enfin, une base de données qui constitue en quelque sorte un prototype du projet d'informatisation conçu par l'auteure. Une annexe présente deux exemples de résultats produits par la base de données.

Dans l'introduction, l'auteure explique que son projet a comme toile de fond la refonte des articles de la tranche alphabétique B-[22], auxquels son modèle de *FEWi* pourrait s'appliquer dans un premier temps.

[22] La refonte des articles de la tranche alphabétique B- est actuellement en cours à l'ATILF, sous la direction de Jean-Paul Chauveau.

Etymon glose explicative
I : formes héréditaires / commentaires
2 critères : - usage ininterrompu
- évolution phonétique régulière

II : emprunts à l'étymon latin / commentaires
Mots savants et semi-savants

Note : distinguer si les formes relèvent de la partie I ou de la partie II pose parfois un problème épineux lorsqu'une forme présente un cas de filiation hybride. Le terme de « double filiation » signifie ici qu'une forme est le fruit, à la fois, de facteurs relevant du domaine I et de facteurs relevant du domaine II. Toutefois, dans la grande majorité des cas ce genre de problème n'est pas susceptible de se poser. On peut donc le considérer comme un phénomène marginal ne remettant pas en cause la pertinence de la hiérarchie sus-présentée.

III : emprunts (importations et substitutions lexèmatiques), calques sur d'autres langues / commentaires

A l'intérieur de chaque partie, les formes dérivées sont, quant à elles, classées par subdivisions sémantiques.

FIGURE 1.2 – Structure hiérarchique des articles du *LEI* (d'après V. Beckert)

1.3.1.1 Analyse structurelle des articles

Sous le titre d'analyse structurelle, l'auteure émet surtout une vision générale, qui se résume à l'observation d'une grande irrégularité dans l'organisation microstructurelle des articles. Virginie Beckert propose de supprimer cette irrégularité en calquant la microstructure irrégulière des articles du FEW sur celle – régulière – du LEI (cf. Aprile 2004, 88 ; 129-131) et fournit le schéma visible à la figure 1.2.

La nécessité de rendre la microstructure du FEW plus régulière est justifiée par le projet d'informatisation du FEW. En effet, si les articles du FEW doivent être introduits dans une base de données – qui est le moyen d'informatisation adopté d'emblée par l'auteure –, il est important de pouvoir délimiter correctement le contenu de chaque champ de cette base de données. Dans le FEW, cette délimitation des champs doit absolument s'effectuer de façon automatique :

> [La délimitation du contenu d'un champ] ne peut être effectuée de façon manuelle du fait de la taille et de la densité du dictionnaire. Il est nécessaire d'automatiser cette tâche et d'examiner dans quelle mesure la structure préexistante est suffisamment manipulable afin de permettre un traitement automatique. On pourra tenter de reconnaître les objets textuels et leur portée grâce à l'écriture de règles (reposant surtout sur des critères typographiques). (Beckert 2003, 11)

Virginie Beckert met ici le doigt sur le problème principal que pose l'informatisation du FEW, à savoir la question de la faisabilité d'un balisage automatique. Un balisage automatique nécessite effectivement l'écriture préalable de règles permettant au programme de reconnaître les éléments à baliser. Or, la question ne reçoit pas une

réponse positive :

> [...] on peut affirmer que la reconnaissance d'informations succinctes ne pose pas de problème majeur. En revanche, du fait de l'irrégularité formelle des articles du FEW dans leurs versions actuelles, la délimitation des champs sur lesquels portent les objets méta textuels devrait, en partie, être opérée manuellement. (Beckert 2003, 12)

Homogénéiser la microstructure des articles selon le modèle du LEI est, d'après l'auteure, la solution qui permettrait une délimitation automatique des champs :

> Toutefois, si la structure des articles est revue en vue de l'informatisation du FEW, selon le modèle du LEI présenté plus haut, la délimitation ne sera plus problématique. Il s'agira de convertir l'ensemble des articles selon le modèle présenté ci-dessus, afin qu'ils obéissent tous à la même structure formelle. Ensuite, on pourra y intercaler des balises qui serviront à identifier et délimiter le champ des différents objets (par exemple, on peut baliser l'ensemble des étymons en plaçant une balise ouvrante <ved> et une balise fermante </ved> respectivement juste avant et juste après chaque étymon). (Beckert 2003, 12)

V. Beckert décrit huit « informations succinctes » qui, étant « toujours présentées de la même façon au niveau typographique [. . .] sont en conséquence clairement repérables et pourront être extraites automatiquement » : les « vedettes », la « glose définissant la vedette », les « sigles géolinguistiques et bibliographiques »[23], les « formes », les « locutions »[24], les « gloses des formes », les « étymons cités à l'intérieur d'articles » et, enfin, les « suffixes en sous-vedette »). Leurs caractéristiques typographiques sont détaillées dans la figure 1.3.

Ces caractéristiques typographiques, associées aux index du FEW (l'index sélectif des formes [ATILF 2003] et l'index des sigles [25]) doivent permettre d'écrire des « règles » permettant de repérer et d'extraire les informations :

> Pour identifier les sigles géolinguistiques ainsi que la forme sur lesquels [sic] ils portent, on pourrait écrire une règle précisant que la chaîne de caractères est un sigle géolinguistique si elle remplit les conditions suivantes :
>
> 1- elle apparaît en romain de taille normale, en non gras et non italiques, et appartient à la nomenclature des sigles.
>
> 2a- elle figure à gauche d'une forme, après un blanc typographique, OU
>
> 2b- elle figure à gauche de la mention id. après un blanc typographique, OU

[23]Cet intitulé regroupe deux réalités au contenu très différent : les indicateurs géohistoriques, situés avant l'unité lexicale, indiquent la langue à laquelle elle appartient, tandis que les sigles bibliographiques mentionnent les sources où ont été trouvées les informations sur l'unité lexicale.

[24]Par « formes », Virginie Beckert désigne les unités lexicales galloromanes ; les « locutions » en constituent un sous-ensemble (représentant les lexèmes plurilexicaux) qu'il ne convient pas, à notre avis, de distinguer ici.

[25]En réalité, il ne s'agit pas vraiment d'un index, mais d'un *Beiheft* cumulatif.

Vedettes	Police : romain. Taille de police légèrement supérieure, gras.
Glose définissant la vedette	Elle se trouve directement à droite de l'étymon après un blanc typographique (sauf si c'est un étymon non-latin, auquel cas la langue est précisée entre parenthèses entre la vedette et son définissant), en romain, même taille de police que la vedette, c'est-à-dire légèrement supérieure au reste de l'article, non gras. Les gloses ne font pas partie d'une nomenclature.
Sigles géolinguistiques et bibliographiques	Ce sont pour la plupart des abréviations. Elles apparaissent en romain, en police de taille normale, non gras, non italiques. Ils font partie d'une nomenclature et les index sont actuellement en cours de vérification et de mise à jour pour une édition révisée et augmentée cette nomenclature. Les sigles sont saisis au fur et à mesure des vérifications avant d'être indexés et balisés. Ils figurent dans les articles en caractères de taille normale, non gras, non italiques. Ils sont systématiquement antéposés pour les sigles géolinguistiques, et post-posés pour les sigles bibliographiques.
Formes	Les formes apparaissent dans les parties I , II ou III d'un article. Elles sont en romain, de taille normale, en italiques. Une partie d'entre elles figure d'ores et déjà dans une nomenclature balisée. Les formes provenant de sources orales et n'étant pas attestées à l'écrit apparaissent sous leur représentation phonétique, en caractères de taille normale, italiques, espacés.
Locutions	elles apparaissent en romain, en italiques.
Gloses des formes	elles apparaissent entre guillemets après la forme qu'elles définissent, en romain, de taille normale.
Etymons cités à l'intérieur d'articles	Les étymons apparaissent en vedette d'article mais peuvent également être mentionnés à l'intérieur d'articles ayant un autre étymon pour vedette. Dans ce cas l'étymon cité figure en romain, en petites majuscules, (non italiques et non gras).
Suffixes en sous vedette	Ils apparaissent en sous vedettes de l'étymon auxquels ils se rapportent. Ils sont précédés du signe + ou X, d'un blanc typographique, d'un trait d'union, le suffixe est donné sous sa forme latine en romain, petites majuscules, (non italiques et non gras).

FIGURE 1.3 – Les objets et leurs caractéristiques typographiques respectives (d'après V. Beckert)

2c- elle figure à gauche d'une définition, après un blanc typographique.

(Beckert 2003, 15)

Virginie Beckert conclut que « c'est ce type d'opération qui va permettre d'identifier la portée des champs des différents objets méta textuels et d'extraire correctement les séquences d'articles selon les différentes interrogations ».

1.3.1.2 Propositions d'interrogations

La base de données imaginée par l'auteure est dotée d'un système d'interrogation qui permette à l'utilisateur de formuler efficacement une question (« requête ») et de recevoir une réponse pertinente (« résultats de la requête »). Le résultat se présente à l'utilisateur sous la forme d'une liste d'articles ou de « séquences d'articles ». Par cette dernière expression, l'auteure entend des extraits d'articles pertinents, prédéfinis en fonction du type de requête.

La base de données proposée par l'auteure comporte trois types de données : (1) les articles du FEW (homogénéisés selon la structure du LEI), (2) un assortiment d'index (appelés aussi *tables*), (3) des séquences d'articles prédéfinies en fonction des types de recherche. Le tout est mis en relation par des liens hypertextes. Chaque entrée d'un index renvoie à un ensemble de séquences (résultats), ce résultat étant dans certains cas (notamment lorsque l'entrée est un étymon) l'article complet.

> La création d'une base de données constituée par les articles étymologiques du FEW [...] pourra s'accompagner de la création d'index et de liens hypertextes permettant d'aligner (i.e. d'associer) les termes des différents index – index des étymons, index des formes – et des types prédéfinis d'information (par exemple : la liste des formes apparentées (héréditaires et emprunts) ; les commentaires correspondant à ces formes ; la liste des étymons mentionnant le lexème recherché dans leur article etc.). L'extraction des éléments de réponse vise à localiser (atteindre), manipuler (aligner) et à faire ressortir les séquences constituant des éléments de réponse à la requête. La qualité d'un tel système dépend de la pré-définition, par les auteurs des articles étymologiques du FEW, des séquences d'articles pertinentes selon les différents types de recherche. (Beckert 2003, 16-17)

Le nœud du problème consiste donc à définir les différents types de recherche souhaités et les séquences (ou extractions) d'articles attendues comme résultat. Il faut en outre spécifier, via l'écriture de règles, de quelle façon le programme peut repérer ces séquences de façon pertinente.

Deux grands types de recherche sont envisagés par l'auteure : la recherche simple d'un lexème (un étymon ou une unité lexicale galloromane) et la recherche complexe, définie comme une combinaison de divers critères d'interrogation (par exemple « les étymons comprenant la chaîne de caractères <elet> en position finale » ou encore « les emprunts à l'anglais antérieurs à 1066 » [Beckert 2003, 20]).

Dans le cas d'une recherche simple, l'interface propose à l'utilisateur d'entrer le lexème (étymon ou unité lexicale galloromane) qu'il recherche. Le programme parcourt la table des étymons ou des unités galloromanes, en extrait le lexème et le présente à l'utilisateur, accompagné si la requête concerne un étymon, de l'article entier

Champ 1	sigle géolinguistique / forme / définition
Champ 2	type de descendance : forme héréditaire de l'étymon x (partie I de l'article où apparaît cette forme) OU emprunt à y (partie II) OU emprunt à z (partie III)
Champ 3	le lecteur pourrait choisir de visualiser l'article complet de l'étymon en cliquant sur un bouton VISUALISER L'ARTICLE ENTIER

Par exemple, pour la forme *esse*, sous classée sous *axis* :

aflandr. ***Esse*** « essieu »
forme empruntée du latin *axis* « essieu » VISUALISER L'ARTICLE ENTIER

FIGURE 1.4 – Exemple du résultat de la recherche d'une forme (d'après V. Beckert)

restructuré et, si la recherche concerne une unité galloromane, de son sigle géolinguistique et de sa définition, ainsi que du type de transmission (forme héréditaire ou emprunt ; appelé ici « type de descendance »), avec un lien vers l'article entier (voir figure 1.4 ; cf. Beckert 2003).

En ce qui concerne les recherches complexes, Virginie Beckert envisage onze possibilités de recherche sur des éléments particuliers : préfixe, suffixe, « étiquette » (sigles géolinguistiques et bibliographiques, datations), mot composé, dérivé, calque, emprunt, chaîne de caractères, locution, sens, formes comparées non galloromanes.

L'auteure fournit pour chacun de ces types de recherche une définition plus ou moins précise de la méthode de recherche envisagée et du résultat qui devrait être fourni à l'utilisateur. La recherche d'un étymon comportant un certain préfixe s'effectue par exemple selon la règle suivante : le préfixe recherché doit figurer dans la nomenclature des préfixes (préalablement constituée), apparaître en position initiale d'un étymon figurant au nombre des étymons recensés dans les index, comporter un blanc à gauche en début de phrase ou de paragraphe[26] et comporter une chaîne de caractères à droite. Cette règle doit permettre à un automate parcourant le FEW numérisé de reconnaître le préfixe et d'extraire l'étymon dans lequel il apparaît. En sélectionnant le préfixe dans l'index des préfixes latins ou en l'entrant manuellement, l'utilisateur obtient la liste des étymons le comportant. Chaque entrée de cette liste comporte un lien hypertexte vers l'article ouvert par l'étymon.

La majorité des règles fournies par l'auteure ont pour objectif la constitution d'un index (des composés, des calques etc.), qui sert ensuite de point de départ à l'utilisateur. En cliquant sur un élément d'un index, l'utilisateur reçoit pour résultat des « séquences d'article » (liste d'étymons, parties d'article etc.) à partir desquelles il est renvoyé vers les articles complets du FEW.

Tous les éléments du FEW ne se prêtent pas à une recherche automatisée. Un groupe de quatre éléments sont définis comme impossibles à rechercher de façon informatique dans l'état actuel du FEW :

[26]Il nous semble que cette condition n'est pas toujours vérifiée, puisqu'un préfixé peut également apparaître dans le commentaire final, à n'importe quel endroit, donc pas nécessairement en début de phrase ou de paragraphe.

1. Recherche d'un mot (semi-)savant : ces mots ne possédant aucun marquage particulier, il est impossible de créer une règle permettant de les extraire automatiquement.
2. Recherche d'un synonyme : cette recherche « nécessiterait la création préalable d'une base de données comportant toutes les relations de synonymie existant entre étymons et formes, si toutefois il est correct de parler de synonymie quand on réfère à des lexèmes diachroniquement très éloignés » (Beckert 2003, 39).
3. Recherche de termes exprimant un concept : contrairement à la recherche d'un sens, qui utilise les gloses, il s'agirait ici d'utiliser un index onomasiologique (Virginie Beckert propose d'enrichir le *Begriffssystem* de Hallig et von Wartburg 1952) afin d'atteindre des lexèmes exprimant un concept déterminé.
4. Recherche de locutions dialectales : cette recherche nécessiterait que chaque locution dialectale soit associée par les rédacteurs des articles aux formes standard (du français moderne) correspondantes. Par exemple, une locution dialectale comportant *bô* serait associée à *bois*, ce qui permettrait de l'atteindre en recherchant la forme *bois*.

1.3.1.3 Prototype de base de données

Dans la dernière partie de son mémoire, Virginie Beckert présente une maquette de base de données, élaborée sous Access, qui permet d'effectuer un certain nombre d'interrogations (telles que celles définies précédemment) dans un corpus de sept articles, appartenant à la refonte de la tranche alphabétique A-. Ces sept articles (AVA, AVIA, AVIUS, AVIOLA, AVIOLUS, AVITINUS et AVUNCULUS) ont été préalablement convertis selon la structure du LEI.

La base de données comporte une seule table, intitulée « Structure article étymologique », qui contient les articles dans leur intégralité. Ces articles sont découpés en différentes portions, selon des champs prédéfinis qui correspondent à différentes parties de la microstructure. S.v. AVUNCULUS, par exemple, se trouvent successivement les champs suivants :

– Vedette,
– Définissant vedette,
– Commentaires étym,
– Formes partie 1,
– Sens secondaires 1,
– Préfixés 1,
– Suffixés 1,
– Composés 1,
– Commentaires 1,
– Locutions 1,
– Confixés 1,
– Dérivés 1,

– avec agglutination [sic],
– Croisements 1,
– Formes enfantines,
– Formes partie 2,
– Composés 2,
– Préfixés 2,
– Suffixés 2,
– Commentaires 2,
– Formes partie 3,
– Commentaires 3,
– Rédacteur,
– Sources biblio.

La base contient également une série de requêtes prédéfinies dont voici les intitulés tels qu'ils sont formulés par l'auteure (Beckert 2003, 47-53) :

– « atteindre tous les articles rédigés par un auteur » ;
– « atteindre l'article de l'étymon avunculus » ;
– « obtenir tous les articles des étymons comprenant la chaîne de caractères <vi> » ;
– « atteindre les formes anglo-normandes classées sous les étymons terminant [sic] par <iolus> » ;
– « rechercher les préfixés en [BIS- +] dans l'ensemble de la table » ;
– « obtenir les composés avec [BELLUS +] dans l'ensemble de la table » ;
– « obtenir les préfixés avec [RETRO- +] dans l'ensemble de la table » ;
– « sens secondaires des étymons aviola, aviolus et avunculus » ;
– « estimer la concurrence des formes héritées du latin ava avec les formes 1 des autres étymons en ancien béarnais » (les « formes 1 » désignant les formes qui descendent d'un étymon par voie héréditaire).

S'ajoutent encore les requêtes prédéfinies consistant à rechercher un sigle (« Québec », « GCoinci », « Ac 1694 », « Boiste 1803 », « Palsgrave », « all. », « Borneo », « aost. »), une forme en *-eul* ou *-ol*, et une date : « En cherchant les chaînes de caractères « 16e » et « 15** » (les astérisques représentent ici des caractères numériques), on est à même de retrouver les attestations du 16ème siècle » (Beckert 2003, 53).

Ces requêtes s'apparentent toutes à une utilisation simple d'Access, qui consiste à effectuer dans la table « Structure article étymologique » la sélection d'un ou de plusieurs champs et ensuite à y rechercher une chaîne de caractères. Voici par exemple la démarche à suivre pour rechercher les préfixés en BIS- :

> En mode feuille de données, Cliquer sur Requête dans la fenêtre de données principale, puis sur Nouveau. Dans le champ de la première colonne, double-cliquer sur le champ Préfixés 1 et inscrire Comme « [BIS- +] », dans le champ Critères. Dans la seconde colonne, double-cliquer sur le champ Préfixés 2, et inscrire Comme « [BIS- +] » dans le champ Ou. Répéter l'opération pour le champ Préfixés 3. Cliquer sur l'icône ! pour exécuter la requête. Localiser les [BIS- +] à l'aide de la commande Rechercher (CTRL + F) dans le champ retourné par la requête. (Beckert 2003, 49)

Toutes les requêtes fonctionnent de la même façon : elles renvoient à de larges portions d'articles (correspondant aux champs définis par l'utilisateur), dans lesquelles l'utilisateur doit lui-même repérer ce qu'il cherche, au moyen de la fonction « CTRL + F » qui met en surbrillance, une par une, les occurrences de la chaîne de caractères demandée (sachant que ce type de méthode peut contenir beaucoup de « bruit », c'est-à-dire des résultats qui ne correspondent pas toujours aux informations recherchées).

1.3.2 Analyse critique

Dèss le premier chapitre de son mémoire, Virginie Beckert repère le problème principal posé par une informatisation du FEW : vu la taille du dictionnaire, la reconnaissance des informations, et surtout des relations qu'elles entretiennent entre elles, doit absolument s'effectuer de façon automatique. Or, cette reconnaissance est problématique dans la mesure où la microstructure des articles est assez souple ; nous ajouterons à cela que l'infrastructure elle-même, soumise pourtant à des règles strictes, pose problème à cause de la présence massive d'implicite (→ 2.6).

La solution « miracle » que trouve Virginie Beckert consiste à uniformiser la microstructure des articles du FEW, selon le modèle du LEI. Cette proposition nous paraît irréaliste. La réécriture des articles du FEW peut effectivement s'envisager dans le cadre de la refonte des articles du premier volume, puisque ces articles (du moins ceux en B- à l'heure actuelle) sont en cours de rédaction. En revanche, il paraît tout à fait impossible de réécrire la totalité des articles du FEW sur le modèle du LEI – au mieux, on pourrait envisager de convertir les articles au moyen d'un programme informatique, mais il faut encore étudier la faisabilité d'une telle opération. Penser, comme Virginie Beckert, qu'il suffit de réécrire les articles serait aller un peu vite en besogne.

Le deuxième chapitre du mémoire confirme le cadre restreint dans lequel se placent les réflexions de l'auteure. Les recherches concernant les locutions, les composés ou les calques ne peuvent être envisagées, dans l'état proposé par Virginie Beckert, que dans les articles de la refonte de la lettre B, qui contiennent des marqueurs explicites pour ces types d'information.

Les recherches envisagées dans ce chapitre suivent toutes la même démarche : elle consiste à repérer les informations et à constituer un index, dont chaque entrée renvoie par lien hypertexte aux articles d'où elle provient.

Il est étonnant de constater que dans le troisième chapitre du mémoire, qui présente la maquette de FEW informatisé réalisée par Virginie Beckert, il n'est plus du tout question de ces index. Toutes les requêtes de la maquette se limitent à la recherche d'une chaîne de caractères (entrée par l'utilisateur) au sein d'un article ou d'une partie d'article (également définie par l'utilisateur). Il n'est plus question ici de champs

correspondant aux objets définis précédemment (tels que l'étymon, les formes, les dérivés etc.) : les seuls champs définis sont ceux de la microstructure, correspondant aux paragraphes d'un article.

La conséquence est l'impossibilité d'obtenir des résultats précis : toutes les requêtes sans exception renvoient à de larges portions de texte, dans lesquelles l'utilisateur doit lui-même repérer le phénomène qu'il cherche. Cette opération s'effectue de manière classique par l'exécution de la commande « rechercher », qui repère dans le texte au cas par cas les occurrences de la chaîne de caractères entrée par l'utilisateur. En somme, la base de données n'est pas exploitée et ne sert pas à grand-chose ; tout au plus permet-elle de choisir l'article ou la portion d'article dans laquelle effectuer la recherche. La meilleure preuve de cette inexploitation de la base est qu'elle ne possède ici qu'une seule table (constituée de l'ensemble des articles), alors que l'intérêt d'une base de données est justement de permettre des liens entre différentes tables.

Nous nous doutons bien que cet état de simplicité de la maquette est dû au contexte temporel très restreint (un stage de quelques mois) dans lequel Virginie Beckert a travaillé. On peut aussi se demander, avec l'auteure, si Access était la base de données la plus adéquate pour ce projet. Virginie Beckert dit elle-même qu'« une maquette de FEWi est un projet qui sort un peu du cadre habituel d'utilisation d'Access. Certains types de requêtes fonctionnent très bien mais d'autres posent problème »[27] (Beckert 2003, 54). Il existe peut-être sur le marché des bases de données plus performantes qu'Access, qui pourraient convenir davantage, si tant est que l'on considère le passage par une base de données comme utile.

1.3.3 Enseignements à tirer de ce travail

Après l'analyse du travail effectué par Virginie Beckert, une constatation s'impose comme une évidence : la question de l'informatisation du FEW représente un problème trop complexe pour un mémoire de DESS. Il ne pouvait être résolu en trois mois de stage, par une étudiante non formée en linguistique historique et non initiée aux structures du FEW. Toutefois, les critiques précédentes mises à part, nous pouvons retenir de ce travail, pour l'informatisation complète du FEW, plusieurs informations importantes.

La première concerne ce que l'auteure appelle les « informations succinctes », dont les caractéristiques typographiques permettent un balisage automatique : vedettes, gloses, sigles, lexèmes, définitions, étymons et suffixes. Virginie Beckert détaille avec précision ces caractéristiques typographiques, que nous pouvons reprendre et affiner afin qu'elles correspondent non pas à l'état de la refonte, mais à la totalité du FEW.

Un deuxième résultat de recherche à utiliser concerne les différents types d'interrogations proposés par l'auteure. En effet, tout projet informatique présuppose une définition précise des besoins des utilisateurs. Dans notre cas, ces besoins se concentrent évidemment dans les possibilités de recherche, prioritaires pour le choix des solutions informatiques. Les types de recherche proposés par Virginie Beckert sont donc à conserver, à condition qu'ils correspondent effectivement aux demandes des utilisateurs potentiels. Il restera également à vérifier que ces recherches soient techniquement

[27] « En effet, ce type de base de données ne sait pas renvoyer des noms de champs correspondant à des conditions dans une requête. Il faut donc passer par du code pour essayer de retrouver l'information, ce qui est plus long et surtout plus difficile pour un novice» (Beckert 2003, 54).

réalisables : cette vérification nécessitera sans doute de profondes modifications quant aux moyens et outils à employer, notamment si l'on opte pour une autre solution qu'une base de données sous Access.

Tout en gardant cette réserve quant aux moyens techniques, nous pouvons également garder en mémoire les idées émises par Virginie Beckert quant à l'interface de recherche, notamment en ce qui concerne les critères à combiner pour permettre à l'utilisateur de restreindre et d'affiner les résultats.

Enfin, certains problèmes rencontrés par l'auteure ne sont pas dus, à notre avis, à l'utilisation d'Access, mais au FEW lui-même. Virginie Beckert prévient notamment les éventuels utilisateurs de sa base de données que la création d'enregistrements dans la table est fastidieuse : « Les utilisateurs qui voudraient s'atteler à cette tâche doivent savoir qu'un temps considérable est nécessaire afin de manipuler le texte brut des articles » (Beckert 2003, 45). Les problèmes concernent notamment :

– la mise en page du texte (les colonnes provenant du FEW scanné comportant de nombreux sauts de ligne, qui doivent être supprimés manuellement un à un avant l'insertion du texte dans les tables) ;

– les caractères spéciaux et notamment les « symboles phonétiques », non reconnus par Access ;

– l'intégration manuelle, dans le corps de l'article, des notes situés à la fin de celui-ci.

Comme nous l'avons déjà souligné auparavant (→ 1.2.4), les problèmes techniques de numérisation et de reconnaissance des caractères seront à résoudre d'une façon ou d'une autre, avant de commencer l'opération d'informatisation du FEW. L'idéal serait de pouvoir travailler directement sur un document propre, fidèle à l'original. Le traitement des notes sera à prendre en compte en même temps que celui des autres informations dans l'étude des possibilités de balisage automatique.

1.4 Le balisage des articles en cours de rédaction

1.4.1 Contexte historique

En 1993, lors du transfert de l'équipe du FEW de Bâle à Nancy, le Conseil scientifique du FEW a assigné à l'équipe de rédaction un programme d'urgence à réaliser de façon prioritaire. Il s'agissait d'une part de terminer la partie du FEW consacrée aux matériaux d'origine inconnue ou discutée (volumes 21-23) : cette partie s'est achevée en 2001, avec la publication du fascicule 160, qui contient une table des matières et un index des concepts de cette partie non étymologique du FEW, dus à Yan Greub. D'autre part, von Wartburg avait entrepris à la fin de sa vie une refonte du tome premier (A-B), rédigé dans les années 1920 (1922–1928). Cette refonte s'imposait pour donner une unité à l'ensemble, les principes organisateurs de l'œuvre ayant évolué au cours de son développement. La publication en 2002 du fascicule 161/162 a terminé la refonte des étymons latins, grecs et préromans à initiale A- (volumes 24 et 25) et l'a même complétée par 70 pages de *corrigenda* à ces deux volumes, dus à Jean-Paul Chauveau. Par

cette publication, l'équipe a achevé la réalisation du programme à moyen terme qui lui avait été assigné dix ans auparavant.

Après l'achèvement de ce programme d'urgence, l'équipe de rédaction du FEW a logiquement continué sur sa lancée en entamant la refonte de la lettre B, qui est située dans le même premier volume du dictionnaire et est sujette aux mêmes révisions. Toutefois, contrairement à la refonte de la tranche alphabétique A-, celle de la lettre B ne suit pas l'ordre alphabétique, mais s'effectue de façon sélective, par tranches d'étymons. Selon les dires de Jean-Paul Chauveau, ce mode de travail a des répercussions sur la publication des articles :

> La suite du travail va de soi : la refonte de la lettre B, dont la première version souffre des mêmes imperfections que celle de la lettre A. Mais, pour des raisons administratives et organisationnelles qui nous sont imposées par nos directions, cette refonte ne pourra suivre l'ordre alphabétique. De ce fait la publication sur papier est, au moins provisoirement, impossible. S'est donc imposée la mise à disposition du public scientifique dès l'immédiat sur un site informatique, au sein de l'ensemble de bases de données de notre laboratoire, dont les plus connues sont la base textuelle *Frantext* et le *Trésor de la langue française informatisé*. (Chauveau 2006, 33)

Une publication électronique ouvre des perspectives nouvelles sur l'accessibilité du FEW. Divers moyens informatiques, que nous détaillons plus loin (→ 1.5), ont été envisagés pour faciliter la consultation de l'ouvrage. Le premier est le balisage des articles, qui permet de rechercher rapidement une information et d'interroger le dictionnaire de façon transversale. Anne-Christelle Matthey, rédactrice au FEW, et Gilles Souvay, ingénieur informaticien, ont été chargés de ce dossier et ont réalisé un prototype de balisage XML pour les articles de la refonte en cours. Le fait que ce balisage ait ensuite été abandonné (au profit d'un projet de rédaction modulaire, cf. Matthey et Nissille 2010) ne le rend pas moins intéressant pour notre étude.

1.4.2 Différences structurelles avec la rétroconversion du FEW

En comparaison avec le problème qui nous occupe, l'informatisation complète du FEW, le balisage des nouveaux articles présente trois différences inhérentes.

La première concerne la place de l'informatique dans l'élaboration du produit lexicographique. L'informatisation complète du FEW s'inscrit dans un processus de rétroconversion, qui consiste à convertir en format électronique un dictionnaire élaboré primitivement sur papier. Le format papier précède donc le format électronique : le projet s'insère dans une logique de *lexicographie informatisée*. La rédaction sous XMetal des articles refondus de la tranche alphabétique B-, en revanche, relève de la démarche opposée, puisque les articles de la refonte seront directement saisis dans le format électronique et respecteront les règles et contraintes de ce format. On parlera de *lexicographie informatique* (cf. Martin 2001). Le balisage conçu dans cette optique peut, par définition, être plus restreint qu'un balisage devant prendre en compte les divers formats et les incohérences presque inévitables d'un dictionnaire déjà existant. Cette divergence entre les deux processus s'aggrave lorsque le dictionnaire papier a été rédigé en plusieurs étapes sur une longue période (en l'occurrence, quatre-vingts ans) et

présente des traces tangibles de changements de méthode[28].

La seconde remarque, qui dérive de la première, concerne la différence de corpus entre ces articles en cours de rédaction et les articles antérieurs du FEW. La particularité majeure des articles de la refonte de la lettre B consiste, nous l'avons dit, à être rédigés dans la perspective d'un traitement informatique immédiat. Conscients dès le départ que ces articles pourraient recevoir un balisage, les rédacteurs ont repensé leur écriture afin de mettre en évidence les informations susceptibles de faire l'objet d'interrogations. Il en résulte une structuration plus codifiée, avec l'obligation de respecter les grandes articulations de l'article et notamment la présentation successive des différentes issues étymologiques (en chiffres romains, → 1.4.3.2). Il en résulte également une explicitation de certaines informations :

– les titres de paragraphe et leur marquage alpha-numérique apparaissent désormais de façon explicite et dans le détail ;

– les affixés (préfixés et suffixés) font l'objet d'une présentation typographique particulière ;

– les principes explicatifs de l'évolution sémantique (métaphore, métonymie, etc.) sont signalés explicitement ;

– les locutions sont prises en compte de façon systématique dans des sous-paragraphes marqués comme tels ;

– les emprunts et, de façon générale, les références aux autres langues sont également explicités, de façon à permettre une recherche sur tous les emprunts à l'anglais, par exemple.

Tous ces éléments n'accédaient pas, ou en tout cas avec moins de transparence, aux structures de surface des articles précédents du FEW. Cette différence de corpus (l'un étant structuré de façon plus détaillée et plus explicite que l'autre) est une deuxième particularité à prendre en compte dans notre examen du balisage de ces articles.

La troisième remarque concerne les conditions de publication des articles de la refonte. À partir du moment où le balisage est conçu pour des articles destinés à être publiés et consultés séparément, il est évident que le traitement des niveaux structurels supérieurs – l'organisation des articles au sein du dictionnaire – se trouve dépourvu de sens. Les concepteurs du balisage ont donc pu laisser de côté, provisoirement en tout cas, ces questions de supra- et de macrostructure non pertinentes dans le contexte qui était le leur. Ils ont volontairement concentré leur attention sur le nœud du problème : le balisage des observables ou « unités minimales de traitement » (Büchi 1996, 116), qui constituent l'infrastructure du FEW (→ 1.4.3.6)[29].

En tenant compte de ces trois divergences entre le projet d'informatisation complète et celui du balisage des nouveaux articles, le second peut fournir une base de réflexion intéressante pour le premier.

[28] Cf. Büchi 1996, 3-4.

[29] Les termes métalexicographiques que nous utilisons pour rendre compte de la structuration du FEW sont repris de Büchi 1996 (voir aussi Rey-Debove 1971) et expliqués dans la suite de ce chapitre.

1.4.3 Analyse du balisage

La DTD (*Document Type Definition*) élaborée par Anne-Christelle Matthey et Gilles Souvay pour la refonte de la tranche alphabétique B- du volume 1 du FEW se trouve en annexes (→ A), accompagnée de quelques explications destinées à faciliter sa lecture (voir aussi Rouleux 2005). Nous proposons ci-dessous un résumé de la structuration proposée, depuis la macrostructure du FEW jusqu'au plus bas niveau de l'infrastructure.

1.4.3.1 Traitement de la macrostructure

L'élément structurel le plus haut que prévoit la DTD de la refonte est l'élément « FEW ». Il englobe le dictionnaire dans sa totalité : c'est ce qu'on appelle l'élément racine du document.

Cet élément « FEW » est directement et uniquement composé d'éléments « Article », destinés à baliser chaque article séparément. Cela signifie que le balisage ne prend pas en compte les subdivisions superstructurelles que sont d'une part la partie étymologique (vol. 1-20 et 24-25), d'autre part les matériaux d'origine inconnue ou incertaine (vol. 21-23) (cf. Büchi 1996, 9-22) ni ce qui relève de la macrostructure (l'organisation des articles à l'intérieur de ces deux parties). Il ne traite pas non plus de ce qui relève de l'épistructure du FEW (les compléments et corrections à la fin de certains volumes) ou de sa parastructure[30]. Cette absence d'élément informatique (et donc de niveau conceptuel) intermédiaire entre le FEW et l'article s'explique avant tout par les conditions de publication des articles de la refonte (→ 1.4.2).

1.4.3.2 Structuration de l'article en sections

Contrairement à la macrostructure du FEW, qui est ignorée, la microstructure est prise en compte dans le balisage. Par « microstructure », on entend dans le FEW les grandes articulations d'un article, c'est-à-dire sa structuration en divers paragraphes et sous-paragraphes (cf. Büchi 1996, 5-6).

Le balisage des articles de la refonte prévoit qu'un article soit composé de quatre éléments successifs : un élément obligatoire « FormesHereditaires et trois éléments facultatifs « FormesSavantes », « FormesEmpruntees » et « Commentaire ».

Ces quatre éléments correspondent à un maximum de quatre sections distinctes qui structurent, au plus haut niveau, la majorité des articles de la refonte. Les trois premières y sont numérotées en chiffres romains et correspondent à différentes issues étymologiques :

I les formes héréditaires, c'est-à-dire les formes qui proviennent de l'étymon par voie populaire ;

II les formes savantes (emprunts au latin) ;

III les formes empruntées à d'autres langues que le latin.

[30]La parastructure concerne, outre le *Beiheft* et son *Supplément*, les nombreux comptes rendus, les propositions d'ajouts ou de corrections et les articles consacrés au dépistage des étymologies doubles (cf. Büchi 1996, 6).

La quatrième section, le commentaire, ne bénéficie pas d'un marqueur numérique, mais est séparée des autres sections par un simple saut de ligne.

Par rapport aux articles hors refonte, cette structuration demande à être nuancée. Trois paramètres sont critiquables : les conditions d'apparition des éléments, leur nom et leur nombre.

1. *Conditions d'apparition des éléments*. L'obligation, pour les articles de la refonte concernés, de contenir un élément « FormesHereditaires » s'explique par les conditions de publication de ces articles. Hors refonte, cet élément ne peut être obligatoire, car il existe des articles consacrés entièrement à des emprunts. En revanche, le commentaire final est obligatoire dans les articles du FEW (du moins dans la partie étymologique, dont il est question ici) (cf. Büchi 1996, 137).

2. *Nom des éléments*. Le schéma proposé pour la refonte impose une structuration rigide (fondée sur le critère de la transmission), qui est contraire à la philosophie du FEW. Les rédacteurs ont en effet non seulement la possibilité, mais même l'obligation d'adapter la microstructure de chaque article en fonction de l'histoire de la famille traitée. Cette philosophie implique une grande souplesse microstructurelle, qui peut (mais seulement pour de très bonnes raisons) s'écarter du schéma idéal donné ci-dessus (voir par exemple Thibault *in* FEW 25, 978a-981b ; 1011a-1012b ainsi que Büchi 1996, 98-104).

3. *Nombre d'éléments*. Un article du FEW peut dès lors comporter davantage de sections. Büchi 1996, 101 relève notamment un article (FEW 2, 1079b-1080a, CONSTANS) où la numérotation en chiffres romains va jusqu'à « VI ».

En outre, il faut constater dans ce schéma l'absence de deux éléments :

1. La vedette de l'article : un article du FEW est toujours précédé de la mention de l'étymon auquel se rattachent, directement ou indirectement, toutes les formes citées dans l'article. Cet étymon est cité en gras et est généralement accompagné de sa glose (une approximation du sens de base de l'étymon).

2. Les notes de fin d'article : le commentaire est en général suivi de notes, qui constituent le dernier champ (facultatif) de la microstructure du FEW (cf. Büchi 1996, 162). Il sera nécessaire de s'interroger sur le traitement de ces informations un peu particulières, puisqu'elles doivent être reliées à leur source dans l'article.

1.4.3.3 Structuration des sections en paragraphes hiérarchisés

La DTD de la refonte prévoit que les deux premières sections, comprenant les formes héréditaires et les formes savantes, soient structurées en plusieurs éléments hiérarchisés. La troisième section (formes empruntées) et le commentaire final sont modélisés comme un texte continu, qui n'est pas structuré plus finement. Cette différence de traitement est discutable et s'explique par la volonté des auteurs de laisser temporairement de côté l'analyse des deux dernières sections (→ 1.4.2). La catégorie « FormesEmpruntees » peut contenir des paragraphes hiérarchisés et devrait donc bénéficier du même balisage que les deux premières sections de l'article. Seule la section « Commentaire » se présente effectivement comme un texte continu ; toutefois, on aimerait

pouvoir y repérer des éléments particuliers, comme la mention d'étymons, d'unités lexicales ou de suffixes, ce qui justifierait un balisage plus fin de cette section.

Dans le FEW, les paragraphes peuvent être hiérarchisés en six niveaux au maximum, selon le marquage alpha-numérique suivant : I. 1. a. α. a'. α' (cf. Büchi 1996, 101). Les articles de la refonte sont volontairement explicites et très rigoureux en ce qui concerne la numérotation des paragraphes, dans le but de faciliter leur consultation, notamment par moyen informatique (→ 1.4.2). Les articles antérieurs présentent des exceptions à cette règle.

Le balisage des articles de la refonte prévoit un élément intitulé « NiveauHierarchique ». Le marquage alphanumérique présent au début d'un « NiveauHierarchique » est balisé au moyen d'un élément « Numéro ». Cet élément englobe tous les niveaux de l'expression alphanumérique. Par exemple, un paragraphe commençant par les mentions « I. 2. c. β. « arme de poing » »[31] serait balisé comme suit :

> <NiveauHierarchique><NumeroTitre><Numero>I.2.c.beta</Numero><Titre>arme de poing</Titre></NumeroTitre>

Comme le montre l'exemple ci-dessus, chaque niveau hiérarchique est en outre doté d'un titre, qui explicite le critère (sémantique dans l'exemple ci-dessus) selon lequel les formes qui le constituent ont été regroupées. D'après la DTD de la refonte, ce titre de paragraphe est obligatoire. Dans les articles rédigés sous la direction de von Wartburg, ces explicitations du contenu sont, en revanche, rarement formulées. C'est au lecteur d'inférer, par une consultation attentive de chaque paragraphe et du commentaire final, le point commun entre les formes qui y sont regroupées. Il conviendrait donc, hors refonte, que l'élément « Titre » devienne optionnel.

1.4.3.4 Structuration du paragraphe en trois parties

La DTD de la refonte prévoit qu'un « NiveauHierarchique » contienne, après les numéro et titre, un ou plusieurs paragraphes. Chaque paragraphe est composé de trois éléments distincts : ce qui s'appelle « Introduction », un élément intitulé « LexicographieFrancaise » et un élément « Dialectes ». L'introduction est facultative et ne peut apparaître plus d'une fois. Les éléments « LexicographieFrancaise » et « Dialectes » peuvent apparaître de 0 à n fois, à condition qu'au moins l'un des deux soit réalisé.

L'introduction correspond aux quelques mots qui apparaissent parfois au tout début d'un paragraphe pour expliciter son contenu, essentiellement dans le cas de paragraphes sans numérotation. Il s'agit en général de phrases nominales, voire de syntagmes fonctionnant comme des titres : on trouve par exemple, dans l'article BASTUM, les mentions *Locutions nominales*, *Locutions verbales* ou *Par métaphore*.

L'élément « LexicographieFrancaise » correspond aux parties du paragraphe qui concernent les informations non dialectales, c'est-à-dire les unités lexicales précédées (explicitement ou non) de la mention d'une langue (français, occitan, francoprovençal) ou d'un état de langue (y compris les anciens dialectes : ancien picard, moyen français, ...) (cf. Büchi 1996, 125).

L'élément « Dialectes » balise les parties de paragraphe qui présentent des informations dialectales modernes, précédées de la mention d'une localisation dialectale telle

[31] Les exemples fournis dans ce chapitre proviennent de l'article BASTUM (Matthey Anne-Christelle, 2006. BASTUM, version provisoire publiée sur le site internet du FEW (www.atilf.fr/few), Nancy, ATILF).

que « Giv. », « Mée », « Nendaz », « Meuse » etc.

Deux remarques sont à formuler ici. Le choix du nom de l'élément « LexicographieFrançaise » serait à revoir, puisqu'il offre un contenu plus varié que ne le suppose son appellation. Par ailleurs, la distinction effectuée entre cet élément et l'élément « Dialectes » apparaît comme critiquable, car fondamentalement contraire à la vision de l'étymologie telle qu'elle se dégage du FEW. La DTD de la refonte pêche, à ce point de vue, par une tendance à créer des niveaux structurels qui ne sont pas identifiés comme tels dans le discours fewien.

1.4.3.5 Regroupement des unités par variété linguistique

Les éléments « LexicographieFrançaise » et « Dialectes » sont structurés plus finement.

Dans les parties balisées « LexicographieFrancaise », un élément « Langue » regroupe au sein d'un même ensemble les unités qui se rattachent à un même système linguistique — il s'agit surtout d'états de langue — tel que « ancien français », « moyen français », « ancien francoprovençal », « ancien béarnais » etc.

L'élément « Langue » contient obligatoirement en premier lieu un élément « Etat [de langue] », qui entoure la mention de la variété linguistique à laquelle se rattachent toutes les formes regroupées dans l'élément « Langue ». En règle générale, les formes se rattachent à un seul état de langue. Il arrive toutefois que plusieurs états de langue soient cités pour une même liste de formes, d'où la répétition de l'élément « Etat ».

Dans les parties balisées « Dialectes », le même principe de regroupement est appliqué, au moyen d'un élément intitulé « UnDialecte ». La mention du dialecte auquel se rattachent les formes regroupées est identifié par un élément intitulé « Localisation ».

Le choix de regrouper, au sein d'un même ensemble, les unités se rattachant à la même variété linguistique (langue, état de langue ou dialecte) va à l'encontre de la modélisation proposée par Büchi 1996. Éva Buchi considère en effet que la mention d'une langue, d'un état de langue ou d'un dialecte, qui forme l'*étiquette géolinguistique*, est située au même niveau infrastructurel que les informations grammaticales, sémantiques et autres données pour chaque forme, comme nous l'expliquons dans la section suivante.

1.4.3.6 Structuration de l'unité lexicale

L'infrastructure du dictionnaire est composée des observables ou « unités minimales de traitement » (Büchi 1996, 5 ; 98 ; 116-118). Büchi 1996 analyse ces unités comme constituées de « molécules ». Les molécules sont au nombre de huit : l'étiquette géolinguistique (mention d'une langue, d'un état de langue ou d'un dialecte), le signifiant, la catégorie grammaticale, le signifié, la datation, la localisation, la référence bibliographique et les « informations complémentaires » (Büchi 1996, 98 ; 116).

Le balisage des articles de la refonte prévoit que chacune des unités soit balisée par un élément « UneFormeL » dans les parties « LexicographieFrancaise » et par un élément « UneFormeD » dans les parties identifiées comme dialectales. La structuration de ces éléments « UneFormeL » et « UneFormeD » diverge toutefois de l'analyse effectuée par Eva Buchi. En effet, nous avons constaté que, dans le balisage des articles

de la refonte, l'état de langue et la localisation n'étaient pas placés sur le même pied que les autres molécules, mais au niveau structurel supérieur.

Cette divergence s'explique par le principe d'économie suivi par le FEW. Après le sigle géolinguistique (représenté dans le balisage des articles de la refonte par l'élement « Etat [de langue] »), le FEW présente toujours en premier lieu la forme en italique, suivie de la catégorie grammaticale en romain et en abrégé, suivie de la définition entre guillemets, suivie des références[32]. Le FEW suit toutefois un principe d'économie qui consiste, à l'intérieur d'un ensemble infra-phrastique (délimité par une ponctuation forte [point ou point-tiret]), à ne jamais répéter une information déjà fournie tant que celle-ci est toujours valable :

> Quand les unités sont enchaînées pour former un groupe d'attestations terminé au moins par un point, les molécules qui les constituent ne sont pas répétées tant qu'elles restent identiques (Baldinger 1974, 29) ; il s'ensuit qu'elles sont toutes facultatives dans la réalisation concrète. Il n'empêche que certaines d'entre elles sont obligatoires au niveau de la structure profonde : si elles n'apparaissent pas explicitement, elles sont sous-entendues. Il s'agit de l'étiquette linguistique, du signifiant, de la catégorie grammaticale et du signifié. Aussi faut-il distinguer ce que nous appellerons *structure de surface* et *structure profonde*, les deux niveaux étant reliés d'une façon relativement cohérente par des règles d'ellipse. (Büchi 1996, 117)

Les règles d'ellipse mises en évidence par Büchi 1996 concernent toutes les molécules, y compris le sigle géolinguistique.

Dans l'exemple « Afr. *bastun* m. "arme de poing" (Roland ; WaceRouA), *baston* (CourLouis—Cotgr 1611 ; Lac (...) », la DTD de la refonte considère l'état de langue *Afr.* comme une information qui est mise en facteur commun de la liste des formes *bastun* et *baston*, tandis qu'Éva Buchi concluait à son ellipse devant *baston*. La DTD de la refonte prend en considération la structure de surface du FEW, tandis que Büchi 1996 révèle sa structure profonde.

Cette divergence de point de vue a des répercussions non seulement sur le traitement de l'étiquette géolinguistique, mais aussi sur celui du signifiant (l'élément « Forme » de la DTD) et des autres molécules. Nous nous permettons d'introduire à ce niveau une nouvelle notion, celle du « peloton », pour désigner la succession de molécules qui compose une unité minimale de traitement (ici balisée par l'élément « FormesL »). Le contenu de l'élément « UneFormeL » est exprimé dans la DTD de la façon suivante :

```
<!ELEMENT UneFormeL
(Forme, (CodeGrammatical?, Definition?, References),
(((CodeGrammatical, Definition?)|Definition), References)*)>
```

La première partie de cette expression, à savoir `(Forme, (CodeGrammatical?, Definition?, References)`, rend obligatoire la mention d'une forme en tête de peloton, ainsi que la présence d'un élément « References » en queue de peloton. En effet, le rédacteur impliqué dans la refonte B- du FEW doit, pour chaque forme, citer obligatoirement et explicitement une date ou une source (nous verrons plus loin, de façon

[32] Aucun contre-exemple de cet ordre ne nous est connu.

plus précise, quelles informations constituent l'élément « References »). Entre la forme et ses références peuvent apparaître le code grammatical et la définition (situation se présentant lorsque ces deux informations ne sont pas ellipsées, donc lorsqu'elles n'ont pas été citées auparavant pour une unité lexicale occupant une position antérieure dans l'enchaînement).

La seconde partie de cette expression, à savoir `((CodeGrammatical, Definition?) | Definition), References)*` permet de résoudre les cas où d'autres informations suivent, toujours pour la même forme (ce qui correspond à l'ellipse de la forme dans Büchi 1996, 117) :

- un autre code grammatical ;
- une autre définition ;
- un autre code grammatical suivi d'une autre définition.

Dans ces cas, la nouvelle chaîne d'informations est obligatoirement reliée à une référence propre. L'astérisque (*) indique que plusieurs de ces chaînes d'informations peuvent apparaître de façon successive.

1.4.3.7 Traitement des références

Le balisage des articles de la refonte prévoit que les quatre éléments « Etat », « Forme », « CodeGrammatical » et « Definition » soient uniquement constitués de texte. En revanche, l'élément « References », qui concerne des informations de grande importance dans le FEW (la datation et la référence bibliographique), fait l'objet d'un balisage plus fin, exprimé dans la DTD de la façon suivante :

```
<!ELEMENT References
((Date, Date?, Ref) | (Fourchette, Source?) | (Ref))+>
```

Cette expression définit trois formats différents par lesquels le FEW peut citer une référence. Toute combinaison de ces trois cas de figure est permise : ils peuvent apparaître dans n'importe quel ordre, avec n'importe quel nombre d'occurrences. Le signe + contraint à ce qu'un des trois formats apparaisse au moins une fois, afin d'exclure un élément « References » totalement vide.

Les trois formats sont les suivants :

- `(Date, Date?, Ref)` : le premier format commence obligatoirement par un élément « Date », suivi éventuellement d'un second de même nature, suivi enfin d'un élément « Ref ». Nous verrons plus loin à quoi correspond cet élément « Ref » (qu'il ne faut pas confondre avec l'élément « References »).
- `(Fourchette, Source?)` : le deuxième format possible consiste en un élément appelé « Fourchette », qui consigne une durée d'attestation (par exemple 1550 – 1700), suivi facultativement d'un élément « Source ».
- `(Ref)` : le troisième format est constitué uniquement de l'élément « Ref ».

Ces trois formats et les éléments qui les composent servent à rendre compte de l'extraordinaire diversité structurelle avec laquelle apparaissent les références dans le FEW, références dont la présentation est malgré tout strictement codifiée. Il s'agit bien d'une différence de structure et non de contenu. En effet, ces trois formats se réduisent en fin de compte à une série de dates et de sources. L'élément « Fourchette » contient uniquement les éléments « Date » et « Ref », l'élément « Ref » contenant lui-même des éléments « Source » (ou similaires). Dans les références, le FEW présente donc systématiquement des informations de datation (« Date ») et des informations bibliographiques (« Source »). Ce qui varie, c'est la disposition structurelle de ces unités. Voyons ces trois formats de plus près.

Premier format : (Date, Date ?, Ref)

L'élément « Date » balise un texte correspondant à une datation. Celle-ci peut être exprimée dans le FEW sous des formats très variés, comme le montre le recensement effectué par Anne-Christelle Matthey lors de l'élaboration de la DTD :

- dates : 1100, 1879, etc.
- siècles : 12e s. (ou n'importe quels autres siècles)
- fourchettes : 12e—15e s. (ou n'importe quels autres siècles)
- dates approximatives : 1381/1388 (une date entre1381 et 1388)
- parties de siècles : 1er t. 12e s. (pour premier tiers du 12e siècle) / 2e t. 12e s. / 3e t. 12e s. / 1ère m. 12e s. (pour première moitié du 12e siècle) / 2e m. 12e s. / 1er q. 12e s. (pour premier quart du 12e siècle) / 2e q. 12e siècle / 3e q. 12e siècle / 4e q. 12e s. / mil. 12e s. (pour milieu du 12e s.) / déb. 12e s. (pour début du 12e siècle) / fin 12e s.
- compléments : env. 1789 (dans les anciens articles, on peut aussi trouver « ca. » pour 'circa', même sens), dp. 1789 ('depuis'), av. 1100 ('avant'), apr. 1100 ('après'). (Matthey 2004)

Cet élément « Date » peut apparaître une seconde fois, ce qui signifie qu'on peut trouver dans le FEW deux datations qui se suivent. C'est le cas lorsque le dictionnaire cite un *ante quem* et un *post quem* (fourchette) :

> *baston / bâton à feu* "fusil" (1522—1610, Hu ; Fur 1690 ; Rich 1706)
>
> <Date>1522</Date><Date>1610</Date> <Ref><Source>Hu</Source> </Ref>

Comme le montrent l'exemple ci-dessus, les dates sont toujours, dans les nouveaux articles de la refonte, suivies de l'indication d'une source[33] : *Rick*, *Hu* (abréviation de *Huguet*). Cette source est balisée dans l'élément « Ref », défini comme suit :

```
<!ELEMENT Ref
(CommentaireSource?, Source+, Secondaire?)>
```

[33] La présence explicite d'une indication de source constitue la réponse du FEW à l'exigence de vérifiabilité de la démarche scientifique.

L'expression ci-dessus signifie que l'élément « Ref » est obligatoirement constitué d'un élément « Source », qui peut être répété. Il sert à baliser une source bibliographique correspondant aux dates citées auparavant :

> *bastons invasifs et de deffense* "armes offensives et défensives" (1547, DuFail, Propos rustiques, éd. La Borderie 67)
>
> <References> <Date> 1547 </Date> <Ref> <Source> DuFail, Propos rustiques, éd. La Borderie 67 </Source> </Ref> </References>

Avant la mention de la première source peut apparaître un commentaire métalinguistique tiré de ladite source, balisé par l'élément « CommentaireSource » ; après la ou les source(s) peut apparaître une source secondaire, balisée par l'élément « Secondaire »[34]. La source secondaire est celle dans laquelle le rédacteur de l'article a trouvé l'information sur la source primaire. Ainsi, dans l'exemple qui suit, les attestations de *AVigne* et *Chastell* ont été relevées dans la base du *Dictionnaire du moyen français* (*DocDMF*) :

> mfr. *beston* "arme" (AVigne = DocDMF), *baston* "arme quelconque" (Chastell = DocDMF ; Coq ; liég. 1509 - 1621, BTDial 28, 247 ; 31, 45, 104 ; 33, 85 ; 34, 171 ; env. 1500, J. Marot, Voyage de Gênes, éd. Trisolini v. 914 ; Voyage de Venise, éd. Trisolini v. 389), "bâton utilisé comme arme dans certains duels judiciaires, spécialement entre vilains" (Chastell = DocDMF)

Deuxième format : (Fourchette, Source ?)

L'élément « Fourchette » est utilisé lorsque un lexème est attesté durant une période comprise entre deux datations : par exemple 12e—15e s., signifiant « attesté entre le 12ᵉ et le 15ᵉ siècle ». La différence avec la formule « Date, Date, Ref » présentée ci-dessus réside dans le fait que dans l'élément « Fourchette », ces deux dates ne proviennent pas d'une même source. La définition de cet élément permet deux cas de figure :

```
<!ELEMENT Fourchette
  ((Date, Ref) | (Ref, (Date | Ref)))>
```

Dans le premier cas de figure (Date, Ref), la fourchette présente comme premier terme (indiquant le début de la période d'attestation) une date, comme second terme (indiquant la fin de la période) une source (avec toutes les possibilités de l'élément

[34]Du point de vue strict des résultats de la recherche scientifique, la mention de la source secondaire ne serait donc pas nécessaire. Elle s'impose en revanche pour des raisons d'honnêteté intellectuelle : l'attestation dans la source première a été découverte grâce à la source secondaire.

« Ref »), comme dans ce cas :

> mfr. *baston* "arme de fût" (1482—Comm, Gdf ; Gay)
>
> <Fourchette> <Date>1482</Date> <Ref> <Source>Comm</Source> <Secondaire>Gdf</Secondaire> </Ref> </Fourchette>

Dans le second cas (Ref, (Date | Ref)), la fourchette présente comme premier terme une source et comme second terme soit une date, soit une source :

> *baston à feu*(Coq—Widerh 1669 ; Gdf ; GdfC ; Rick ; Li ; J. Marot, Voyage de Venise, éd. Trisolini v. 676)

L'élément « Fourchette », tel qu'il est défini, permet donc les possibilités Date-Source, Source-Date et Source-Source, en excluant la possibilité Date-Date. Cette exclusion est justifiée, puisque ce dernier cas est réservé à la première possibilité de référencement citée ci-dessus (Date, Date ?, Ref). On peut toutefois s'interroger sur la pertinence de cette distinction.

La présence de l'élément facultatif « Source » à droite de l'élément « Fourchette » s'explique dans des cas comme celui-ci, où la source « BTDial 28, 247 ; 31, 45, 104 ; 33, 85 ; 34, 171 » suit la fourchette « 1509–1621 » :

> baston "arme quelconque" (Chastell = DocDMF ; Coq ; liég. 1509–1621, BTDial 28, 247 ; 31, 45, 104 ; 33, 85 ; 34, 171 ; env. 1500, J. Marot, Voyage de Gênes, éd. Trisolini v. 914 ; Voyage de Venise, éd. Trisolini v. 389)
>
> <Fourchette> <Ref> <Source>liég.1509</Source> </Ref> <Date>1621</Date> </Fourchette> <Source>BTDial 28, 247 ; 31, 45, 104 ; 33, 85 ; 34, 171</Source>

Remarquons que cet exemple présente un traitement critiquable de la chaîne « liég. 1509 », réduite à une seule unité, balisée en tant que « Source » (et non pas, comme on pourrait s'y attendre, une source suivie d'une date).

Troisième format : (Ref)

L'élément « Ref » a déjà été expliqué ci-dessus, en tant qu'élément intervenant dans le premier des trois formats (Date, Date ?, Ref). Utilisé ici indépendamment de l'élément « Date », il permet de baliser la mention d'une source non précédée d'une date, comme Roland et WaceRouA dans l'exemple ci-dessous :

> Afr. *bastun* m. "arme de poing" (Roland ; WaceRouA)

Le *Beiheft* (complété, pour les sigles renvoyant à des éditions de textes anciens, par le complément bibliographique du DEAF[35]) renferme la clé de la « traduction » de ces sources en dates : Roland → ca 1100 ; WaceRouA → ca 1170, etc.

Nous avons affirmé plus haut que toute combinaison de ces trois cas de figure était possible. Une forme peut par exemple être suivie d'une fourchette, puis de plusieurs dates avec leurs références. Afin de montrer la complexité de l'élément « References »,

[35] `http://www.deaf-page.de/bibl_neuf.htm`

voici un exemple de référence complète avec son balisage :

> baston "arme quelconque" (Chastell = DocDMF ; Coq ; liég. 1509 - 1621, BTDial 28, 247 ; 31, 45, 104 ; 33, 85 ; 34, 171 ; env. 1500, J. Marot, Voyage de Gênes, éd. Trisolini v. 914 ; Voyage de Venise, éd. Trisolini v. 389)
>
> <References> <Ref> <Source>Chastell</Source> <SecondaireVerifiee="oui"> DocDMF </Secondaire> </Ref> <Ref> <Source>Coq</Source> </Ref> <Fourchette> <Ref> <Source> liég.1509</Source> </Ref> <Date>1621</Date> </Fourchette> <Source>BTDial 28, 247 ; 31, 45, 104 ; 33, 85 ; 34, 171</Source> <Date>env. 500</Date> <Ref> <Source>J. Marot, Voyage de Gênes, éd. Trisolini v. 914</Source> </Ref> <Ref> <Source>Voyage de Venise, éd. Trisolini v. 389</Source> </Ref> </References>

Traitement des références dans les parties dialectales

Dans l'élément « UneFormeD », l'élément « References » n'est pas obligatoire. En effet, pour beaucoup de dialectes, la mention de la localisation (qui, comme on l'a vu, précède celle de la forme) joue en même temps le rôle de référence : les formes localisées « Aire » (= Aire-sur-l'Adour, commune du département des Landes), par exemple, correspondent d'office à la source « Bulletin de la Société de Borda 30, 77 » (donnée dans le *Beiheft*). Il est donc inutile d'ajouter cette dernière comme référence explicite : ce serait une redondance, contraire au principe d'économie qui régit fondamentalement le FEW.

En revanche, le balisage des articles de la refonte prévoit que les unités rattachées à un dialecte puissent être suivies d'un élément facultatif, intitulé « AutresReferencesD ». Cet élément permet de baliser notamment les renvois bibliographiques qui apparaissent à la fin de certains paragraphes (ainsi, dans l'article BIRRUS : « — ALG 1474, 1474*, 656* »). Ces indications ajoutent des références générales sans les associer à un dialecte ou un lexème particulier. Le nom de balise est critiquable, car quelquefois il peut s'agir de renvois bibliographiques concernant le français.

1.4.3.8 L'élément « Derive »

Le balisage des articles de la refonte prévoit un élément facultatif « Derive », qui peut apparaître avant l'élément « Langue » ou l'élément « UnDialecte ».

Cet élément sert à baliser les affixes (préfixes, suffixes ou infixes) qui, dans les articles de la refonte, apparaissent en début de paragraphe, notés entre crochets et en petites capitales. Ces affixes sont mis en facteur commun de la liste des unités les comportant, comme dans l'exemple suivant :

> I. 3. Formes dérivées. [+ -āceu] Béarn. berretàs m. pl. "grands, larges bérets" Palay 1961 ; berretasse f. "(sens métaphorique) grande panse" (tous les deux Palay 1961). [+ -āriu] Apr. beretier m. "bonnetier" (1484, Pans), berretier (1ère m. 13e s., Riquer 1319 = Rn), occit. berretié M, béarn. berretè "celui qui fait, qui vend des berrets", berreté.

1.4.4 Conclusion

Le balisage décrit ci-dessus s'est révélé non applicable en pratique : il ne permettait pas de rendre compte de toutes les situations rencontrées dans les articles de la refonte. Après des essais infructueux sur quelques articles en cours de rédaction, il a finalement été abandonné, au profit d'un projet de rédaction modulaire (cf. Matthey et Nissille 2010).

Cet abandon ne met toutefois pas en cause la totalité de l'analyse qui a été fournie à cette occasion. La question s'est posée de savoir si ce balisage pourrait convenir, en partie au moins, à la totalité du FEW. Son adaptation aux articles hors refonte devrait tenir compte des trois remarques qui ont été formulées en début de chapitre (→ 1.4.2) concernant la place de l'informatique dans le processus, la spécificité du corpus concerné et les conditions de publication des articles.

Les conditions de publication sont responsables des quelques lacunes que nous avons repérées lors de notre analyse. La DTD de la refonte devrait être élargie aux niveaux structurels supérieurs à l'article, volontairement laissés de côté par les auteurs de la DTD : la superstructure, qui est absente, et l'organisation macrostructurelle, qui dans la DTD est réduite à sa plus simple expression. En revanche, les niveaux micro- et infrastructurels ont été développés avec détail et rigueur. Les seuls problèmes de taille que nous ayons relevés concernent la terminologie utilisée (parfois mal choisie, voire contraire à la conception fewienne de l'étymologie) et la modélisation de l'infrastructure, révélatrice d'une divergence conceptuelle avec Büchi 1996. Cette divergence entre deux optiques (l'une à finalité informatique, l'autre métalexicographique) n'est pas banale, car le balisage est censé refléter la structure du document. Dans le cas du FEW, il sera bon de s'interroger sur la meilleure option : a-t-on intérêt à refléter la structure de surface du FEW, ou sa structure profonde ?

Les questions concernant le corpus sont plus délicates. Il faudrait faire en sorte que la DTD des articles de la refonte soit appropriée aux articles antérieurs du FEW, qui n'ont pas été écrits dans l'optique d'une informatisation. Le travail consisterait à comparer les articles de la refonte et les articles antérieurs en dégageant surtout les différences de contenu et les différences de structure. Les premières concernent les noms des balises, qui devraient – après correction des approximations terminologiques de la DTD – sensiblement rester les mêmes pour les deux projets : les nouveaux articles n'ont pas profondément remis en question les types d'information donnés par le FEW. La refonte a surtout donné lieu à un travail d'explicitation des informations cachées, qui touche davantage la structure du document. L'infrastructure des articles, contenant les informations moléculaires, est sans doute moins affectée par ces changements que la microstructure.

Il faudrait enfin examiner l'adéquation de la DTD à un traitement de rétroconversion automatique : il s'agirait de déterminer si tous les éléments prévus dans le balisage pourraient être repérés automatiquement. Cet examen concerne non seulement la structure des éléments, mais aussi leur présentation typographique dans le dictionnaire. Anne-Christelle Matthey avait déjà fait une partie du travail en répertoriant les différents formats de datation, d'états de langue et de citation des sources. Ce relevé, d'une très grande utilité, devrait être vérifié par l'examen attentif des articles antérieurs du FEW et complété par l'étude du contexte dans lequel apparaissent les éléments.

En résumé, le balisage élaboré pour les nouveaux articles n'est pas applicable tel

quel à l'intégralité du dictionnaire, mais il fournit une base de réflexion utile. L'analyse de ce balisage a mis en lumière certaines particularités structurelles du FEW et a montré que plusieurs interprétations étaient possibles de certains éléments, selon qu'elles privilégiaient la structure profonde de l'ouvrage ou sa structure de surface.

1.5 Les autres projets informatiques autour du FEW

En attendant l'informatisation complète du FEW, entreprise trop gigantesque pour être entamée sans étude préalable[36], une série d'autres projets ont été envisagés, qui portent sur une partie du FEW et tentent de donner une réponse utile et rapide à des questions soulevées par la communauté scientifique.

1.5.1 Saisie du fichier onomasiologique

Il existe à l'ATILF, qui est dépositaire des ressources documentaires du FEW, un fichier onomasiologique constitué dans les années 1960 par Walter Lacher et Marianne Müller. Les 30 000 fiches environ de ce fichier recensent les formes contenues dans les 17 premiers volumes du FEW, en les classant selon les concepts du *Begriffssystem* (Hallig et von Wartburg 1952). Un tel index peut se révéler d'une grande utilité :

> Cet index n'est certainement pas parfait, mais tel qu'il est, je sais, pour l'avoir utilisé, qu'il peut rendre d'importants services et qu'il permet d'atteindre des informations plus aisément que par toute autre voie actuellement. Il n'est que de saisir et d'interclasser le contenu de ces quelques milliers de fiches, travail qui va prochainement prendre la suite de l'indexation des formes. (Chauveau 2006)

La saisie de ces fiches a d'ores et déjà été entamée par Nicole Barre, technicienne de classe supérieure à l'ATILF. Cumulé à la parution de l'index sélectif des formes (ATILF 2003), ce projet rendra sans aucun doute de grands services à la communauté scientifique :

> L'indexation des quelques tomes manquants complétée par l'index des concepts des tomes consacrés aux « inconnus » (qui est déjà publié) fournira avec une bonne approximation les données du tableau onomasiologique prévu [par Wartburg]. Combiné à l'« index verborum », cet « index rerum », si l'on peut dire, devrait permettre d'accéder au texte du FEW à la fois par le plan de l'expression et par le plan du contenu. Ce devrait être un excellent instrument de recherche sur l'histoire lexicale du galloroman, mais il devrait également faciliter la recherche proprement étymologique. (Chauveau 2006)

[36] « [...] ces aménagements ne sauraient remplacer l'informatisation totale du FEW, mais celle-ci reste hors de notre portée actuellement » (Chauveau 2006).

1.5.2 Informatisation des étymologies doubles

Ce projet récent consiste en l'élaboration d'une DTD pour le balisage des articles consacrés aux étymologies doubles (Chauveau 2006, 37). Il est actuellement en cours à l'ATILF par Jean-Paul Chauveau et Gilles Souvay.

1.5.3 Version électronique de l'index sélectif des formes

Le directeur du FEW envisage d'élaborer une version informatisée de l'index publié en 2003, qui comprendrait non seulement les formes retenues dans la version papier, mais aussi les lexies complexes, qui n'ont pu être retenues faute de place :

> [L'index du FEW] n'est que la version réduite, pour des raisons de format, de volume et de coût, de l'ensemble des formes qui ont été retenues et qui sont d'ores et déjà mémorisées. L'élaboration d'une version complète informatisée, qui recensera, en plus, toutes les locutions traitées par le FEW (dont le classement n'est pas toujours prévisible), sera entreprise dans la foulée, ce qui ne représentera pas un très gros travail supplémentaire par rapport à la version réduite. Cette version a pour vocation de figurer sur le site du FEW. (...) l'index informatisé devrait permettre au lecteur d'entrer commodément dans le cœur de l'ouvrage, d'y repérer immédiatement les doublets étymologiques et de prendre connaissance de locutions éventuellement insoupçonnées. (Chauveau 2006)

1.5.4 Version électronique du *Beiheft*

La dernière mise à jour du *Beiheft* remonte à 1989. Or, « depuis cette date, quelques milliers de sigles nouveaux ont été forgés pour tenir compte des travaux lexicographiques qui ont été publiés pendant près d'un demi siècle » (Chauveau 2006, 35). La mise au point d'une nouvelle version actualisée du *Beiheft*, conçue comme une base de données informatique, a été entamée par Christian Seidl et Yan Greub en 2002 et a connu un premier achèvement en 2010, avec la publication, sous forme imprimée, du *Complément* (3ème édition du *Beiheft*).

1.5.5 Mise en ligne du manuel d'utilisation du FEW

Dans le même élan que les deux projets précédents, mais avec plus de facilité, il s'agira d'intégrer au site internet du FEW une partie d'un « manuel d'utilisation du FEW » en cours de rédaction par une actuelle et un ancien rédacteurs du FEW (Matthey et Thibault en préparation).

1.5.6 Cartographie informatique

Dans le cadre de la révision du *Beiheft*, Yan Greub avait imaginé un projet de cartographie automatique, qui indiquerait sur une carte le point ou la zone géographique correspondant à une abréviation du *Beiheft*. L'informatisation de ce dernier est un préalable à la réalisation d'une telle carte.

Dans le cadre de l'informatisation complète du FEW, on pourrait imaginer d'aller plus loin encore, en permettant au lecteur de visualiser sur la carte du domaine galloroman la localisation des occurrences citées dans l'article qu'il consulte. Le choix de ces occurrences (par article entier, par paragraphe, par sélection d'unités particulières) permettrait d'avoir une vision géographique précise d'un phénomène particulier et rendrait ainsi de grands services à la recherche.

1.6 Conclusion

L'examen des projets déjà entrepris confirme les craintes exprimées quant à la faisabilité d'un projet d'informatisation du FEW. Plusieurs particularités du FEW apparaissent comme problématiques dans le cadre de son informatisation.

Un premier problème posé par le FEW est la présence de nombreux caractères phonétiques non standards, problème qui a été soulevé par Robert Martin (→ 0) comme un obstacle majeur à l'informatisation de l'ouvrage.

Une deuxième difficulté provient de la masse de données contenues dans le dictionnaire. L'index sélectif des formes, qui ne représente qu'une infime partie des lexèmes analysés dans le dictionnaire, s'étend sur deux tomes. L'étendue du domaine linguistique pris en compte dans le FEW explique la grande diversité des sigles géolinguistiques et l'impossibilité de vérifier leur cohérence : la création de l'index sélectif des formes a notamment permis la découverte de plusieurs sigles erronés, insoupçonnés jusqu'alors. Il faut s'attendre à ce que de telles variations, qu'elles se justifient ou qu'elles soient analysables en termes de « fautes », se présentent également pour d'autres types d'information, par exemple pour les sigles bibliographiques. Cette variété constitue une difficulté majeure dans l'optique d'une reconnaissance automatique des types d'informations, surtout si cette dernière se fonde sur la création d'index, comme Virginie Beckert l'a proposé dans son mémoire de DESS.

L'étude réalisée par Virginie Beckert et la tentative de balisage des articles en cours de rédaction montrent que le discours fewien pose également des problèmes de modélisation. Sa structure ne se laisse pas appréhender facilement : sa microstructure est trop souple, et son infrastructure trop complexe, pour entrer dans un cadre rigide, que celui-ci se présente sous la forme d'une base de données ou d'un balisage XML minutieux. La proposition de réécrire les articles du dictionnaire de façon à harmoniser leur microstructure est peu convaincante, et irréalisable en pratique. Il faut trouver le moyen de modéliser le discours fewien tel qu'il se présente, avec les difficultés qu'il pose.

L'informatisation du FEW apparaît manifestement comme un projet complexe. Il ne peut être résolu sans une réflexion approfondie sur l'objet lexicographique lui-même et sur ses particularités.

Chapitre 2

Le FEW vu par l'utilisateur

2.1 Introduction

Les tentatives qui ont eu lieu dans les années 2003 à 2006, à différents niveaux, pour ouvrir la voie à une informatisation du FEW présentaient des objectifs très différents (→ 1). Néanmoins, elles partageaient un point commun : autant la DTD de la refonte que la base de données Access ou l'index partiel des formes requéraient une compréhension des structures du dictionnaire. L'identification des types d'information présents dans le FEW et de leurs relations constitue une étape préalable et nécessaire à l'informatisation.

Depuis 1996, nous disposons d'une excellente analyse des structures du FEW grâce à la thèse d'Éva Buchi, qui nous sert de référence et de point de départ (Büchi 1996). Cette analyse rend compte de la façon dont le FEW a été *conçu* et *rédigé*. L'analyse des structures du FEW proposée dans ce chapitre et dans le suivant sera quelque peu différente, car notre objectif est de rendre compte de la façon dont le FEW est *perçu* et *utilisé*.

L'attention portée à l'utilisateur d'un dictionnaire est assez récente en lexicographie (cf. Hartmann 2001, 80), mais il est aujourd'hui admis que le concepteur d'un dictionnaire doit d'abord définir précisément le public auquel il s'adresse :

> [...] the most important single piece of advice we can give to anyone embarking on a dictionary project is : know your user. [...] This doesn't imply a superficial concern with 'user-friendliness', but arises from our conviction that the content and design of every aspect of a dictionary must, centrally, take account of who the users will be and what they will use the dictionary for. (Atkins et Rundell 2008, 5)

Nous sommes persuadée que cette remarque s'applique aussi dans un projet d'informatisation d'un dictionnaire existant, dès lors que l'objectif est de résoudre des problèmes d'utilisation : il nous paraît évident qu'il faut avoir une idée de l'identité des utilisateurs du dictionnaire, de ce qu'ils y cherchent, de la façon dont ils le consultent, des difficultés qu'ils rencontrent et des fonctionnalités qu'ils voudraient y trouver.

Dans le but de répondre à ces questions, notre expérience du FEW a été enrichie par des discussions avec plusieurs spécialistes et avec les membres de la rédaction du FEW, ainsi que par les réponses à un questionnaire distribué en 2007, à l'occasion du XXV[e] Congrès International de Linguistique et de Philologie Romanes à Innsbruck, au sein de la communauté internationale des chercheurs en linguistique française et romane. Ce questionnaire avait deux objectifs : tout d'abord obtenir un aperçu des pratiques actuelles d'utilisation du FEW, en déterminant notamment quels types d'informations y sont généralement recherchés ; ensuite connaître les souhaits des utilisateurs dans l'optique d'un FEW informatisé, souhaits qui sont en relation étroite avec les difficultés qu'ils rencontrent dans la consultation de la version imprimée. Les questions à cocher permettaient aux sondés de donner des précisions ou de commenter leur réponse.

Il est à souligner que ce questionnaire a été diffusé à petite échelle et essentiellement à des scientifiques qui utilisent régulièrement et connaissent bien le FEW. Sous sa forme actuelle, en effet, le FEW n'est accessible qu'à un public averti. La conséquence positive de cette situation est que les réponses reçues manifestent une connaissance parfois très profonde et très lucide de ce qu'est réellement le FEW et de ses limites.

Ce chapitre propose une synthèse des résultats de notre enquête. Il a pour but de dégager les bases réflexives qui permettront de construire, dans le chapitre suivant, un modèle du FEW qui puisse répondre aux besoins de la communauté. Nous tenterons tout d'abord de déterminer qui utilise le FEW et dans quel but (→ 2.2), avant de définir comment le FEW est utilisé ou, en d'autres termes, quels sont les comportements de consultation et de lecture de sa version papier (→ 2.3). Les fonctionnalités attendues par rapport à son informatisation seront ensuite prises en compte : les attentes exprimées par les utilisateurs sont en effet révélatrices de la façon dont ils perçoivent l'ouvrage (→ 2.4). Nous montrerons que cette perception est double (→ 2.5). Enfin, nous nous concentrerons sur le problème de l'accessibilité des données et de l'implicite présent dans le discours fewien (→ 2.6).

2.2 Les utilisateurs du FEW

Dans cette thèse, l'« utilisateur du FEW » est souvent présenté de manière dépersonnalisée comme s'il s'agissait d'une entité abstraite. Or, diverses catégories d'utilisateurs consultent le FEW, avec des besoins variés et avec des ressources différentes face à la complexité du discours lexicographique. Avant de tenter de catégoriser les utilisateurs du FEW, il convient par ailleurs de s'interroger sur l'utilité de l'ouvrage.

2.2.1 Utilité du FEW

Si l'on en croit son titre, le FEW est un dictionnaire étymologique du français, ce qui pourrait faire croire qu'il est essentiellement utilisé pour connaître l'étymon des lexèmes de la langue française. En réalité, le titre de l'ouvrage est réducteur (cf. Büchi et Chambon 1995, 947-948). Certes, le FEW sert à *étymologiser* des lexèmes, mais l'étymologie-histoire qu'il pratique le mène à donner davantage d'informations que les autres dictionnaires étymologiques. Une différence importante entre ces dictionnaires et le FEW est que celui-ci présente une étymologie intégrante, c'est-à-dire que l'information étymologique représente le critère organisateur des données, alors que les

autres dictionnaires étymologiques du français intègrent l'étymologie comme information microstructurelle associée à un lexème-vedette. Cette particularité du FEW représente, pour l'utilisateur non spécialiste qui cherche l'étymologie-origine d'un lexème, une difficulté majeure qui incite à privilégier les dictionnaires à étymologie intégrée. Ceci étant dit, le FEW représente la source scientifique de laquelle dépendent ces dictionnaires. Par ailleurs, le domaine couvert par le FEW dépasse la langue française pour embrasser de façon presque exhaustive la totalité des lexèmes du domaine galloroman. Il s'agit donc d'un ouvrage de référence en la matière.

L'étendue du domaine linguistique pris en compte explique que chaque lexème soit associé dans le FEW à une étiquette géolinguistique, qui précise l'état de langue (ancien français, moyen français, français moderne ; ancien gascon etc.) ou le dialecte (lorrain, champennois, ancien picard etc.) auquel il appartient. Le FEW sert donc également à *localiser* un lexème dans un sous-domaine linguistique. Il est important de souligner que l'étymologie et la localisation géolinguistique constituent deux informations *originales* du FEW, c'est-à-dire résultant d'une analyse nouvelle des données effectuée par le rédacteur, contrairement à d'autres informations qui sont reprises telles quelles dans les sources utilisées par le dictionnaire. Une troisième information essentielle donnée systématiquement par le FEW est la datation des lexèmes ou de leurs traits constitutifs (signifiant, catégorie grammaticale, signifié), datation fournie de façon moins explicite dans le programme lexicographique, mais obligatoirement présente (cf. Büchi 1996, 126).

Le FEW sert donc prioritairement à *étymologiser, localiser et dater* un lexème du domaine galloroman. Outre ces trois utilisations qui visent à obtenir des informations sur un lexème vu de façon isolée, l'analyse fournie par le FEW dans le cadre d'une étymologie intégrante permet aussi de *situer un lexème dans un ensemble chronologique et géographique plus vaste* : dans l'histoire d'une famille lexicale galloromane ainsi que, souvent, par rapport aux autres langues romanes. Par situer, on entend notamment expliciter les relations de parenté qu'il entretient avec les lexèmes qui l'entourent. Ces mises en contexte sont également produites directement par l'analyse fewienne et constituent un des principaux intérêts de l'ouvrage.

En dehors de ces utilisations spécifiques au FEW, car en accord avec ses objectifs et son programme lexicographique, des utilisations « indirectes » du FEW sont possibles. Il est par exemple possible de consulter le FEW pour connaître le sens d'un mot, sa graphie, sa forme phonique ou sa catégorie grammaticale. Ces informations sont en effet présentes dans le discours lexicographique. Néanmoins, elles ne constituent pas l'objet de ce discours. Elles dépendent des sources du FEW et sont généralement transmises telles qu'elles sont données par ces sources. L'utilisation du FEW pour accéder à ces informations est donc légitime, mais indirecte, dans le sens où le FEW n'a pas été conçu et pensé dans ce but.

2.2.2 Domaines d'utilisation et catégories d'utilisateurs

La nature des données contenues dans le FEW, à savoir le lexique des parlers du domaine galloroman, explique qu'il soit utilisé en linguistique historique, dans les études concernant le lexique du français et des autres langues ou dialectes du domaine galloroman. Il n'est pas surprenant non plus qu'il soit systématiquement utilisé par les étymologistes des autres langues romanes, ainsi que des langues non romanes. Les premières sont concernées dès lors que le FEW opère une mise en contexte des lexèmes dans des

ensembles linguistiques plus vastes et fournit à cette occasion des parallèles romans (cf. Büchi 1996, 141-145). Les non romanistes (germanistes, orientalistes etc.) sont quant à eux essentiellement intéressés par les étymons appartenant aux langues autres que romanes (volumes 15 à 20 du FEW). De manière plus générale, le FEW constitue une référence pour l'étude historique de toute langue qui a été en contact étroit avec le français.

Parmi les disciplines concernées par l'utilisation du FEW, la dialectologie se trouve au premier plan. Les dialectologues trouvent dans le FEW les matériaux leur permettant soit d'assurer l'étymologie de mots auparavant non expliqués, soit de revoir l'étymologie de mots que l'on pensait connus. Les philologues et éditeurs de textes anciens trouvent dans le FEW une référence en matière d'attestations. L'onomastique n'est pas en reste, puisque de nombreux étymons du FEW sont des noms propres (cf. Büchi 1991 ; 1992). Parmi ces disciplines est également à inclure la lexicographie des aires linguistiques considérées, dont la lexicographie régionale (DRF, DSR, TLFQ) : en effet, même si le FEW ne distingue que depuis récemment de façon explicite les lexèmes régionaux des lexèmes dialectaux, cette information peut être retrouvée via ses sources. Les travaux sur les créoles à base lexicale française requièrent aussi la consultation systématique du FEW. Les utilisations indirectes du FEW expliquent qu'il soit également utilisé, mais de façon moins systématique, dans des disciplines à orientation davantage synchronique, par exemple en sémantique ou en phraséologie.

De ce qui précède, on devine que les utilisateurs du FEW, loin de former un ensemble homogène, disposent de compétences variées ; il est néanmoins certain que ce sont majoritairement des spécialistes en leur domaine, susceptibles dès lors de critiquer les données[1] fournies par le dictionnaire. On pourrait tenter de dessiner une typologie des utilisateurs du FEW selon trois dimensions. Une première dimension concerne le niveau de spécialisation de l'utilisateur dans une des disciplines scientifiques mentionnées ci-dessus. Trois groupes se dessinent : les spécialistes, les étudiants en formation et les amateurs éclairés. Une deuxième distinction peut s'établir entre les utilisateurs critiques du FEW, susceptibles de corriger ou de compléter les données du dictionnaire, et ceux qui consultent l'ouvrage pour en retirer des informations sans mettre en doute son contenu. Enfin, une troisième dimension peut certainement être établie pour distinguer les utilisateurs formés au FEW, qui connaissent la façon dont il doit être utilisé, des utilisateurs débutants ou occasionnels qui ne sont pas initiés aux structures du dictionnaire.

Dans la perspective d'un FEW informatisé, les disciplines et les catégories définies ci-dessus devraient rester similaires, mais l'on peut s'attendre à ce que l'informatisation provoque une extension du nombre d'utilisateurs dans chacune d'entre elles. Il est par exemple indéniable que l'explicitation et la traduction des termes techniques allemands présents dans le FEW aura pour conséquence une utilisation plus importante du FEW au sein du public francophone, qu'il s'agisse d'un public de spécialistes, d'étudiants ou d'amateurs.

[1]Par *donnée*, nous entendons toute information donnée par le FEW, quelle que soit sa nature.

2.3 Itinéraires d'utilisation actuels

Une analyse des comportements d'utilisation du FEW doit distinguer deux modes d'utilisation : d'une part, la *consultation* du FEW, opération consistant à repérer dans le dictionnaire l'endroit où se trouve l'information que l'on recherche ; d'autre part, la *lecture* du FEW, opération consistant à s'approprier de façon complète l'information recherchée ainsi que l'analyse qu'offre le FEW en rapport avec cette information.

2.3.1 Comportements de consultation

La majorité des consultations du FEW ont pour point de départ un lexème connu de l'utilisateur, dont il cherche le plus souvent l'étymologie, la datation (notamment la première attestation) ou la localisation au sein d'une aire géolinguistique. L'utilisation du FEW pour connaître le sens d'un lexème semble également très fréquente, même si quelques réponses au questionnaire signalent – avec raison – que l'aspect sémantique n'est pas toujours précis dans le FEW.

Cherchez-vous fréquemment dans le FEW

- l'étymologie d'un lexème particulier (« adresse FEW ») : 31 réponses positives (97 %)
- la date d'apparition d'un lexème particulier : 29 réponses positives (91 %)
- le sens d'un lexème particulier : 24 réponses positives (75 %)

Le point d'entrée dans le dictionnaire est donc très majoritairement le lexème[2]. Ce type de consultation qui part du lexème est celui qui serait attendu dans un dictionnaire à étymologie intégrée, où les lexèmes constituent les lemmes. Dans le FEW, où les entrées sont des étymons, il revêt un caractère un peu particulier. Pour trouver le lexème dans le dictionnaire, l'utilisateur doit consulter les index du FEW[3] ou un dictionnaire (notamment étymologique : TL, TLF etc.) qui a pour entrée les lexèmes et qui donne leur *adresse FEW*. L'adresse FEW indique l'endroit précis du FEW où se trouve une information (notamment l'article, le lexème ou l'étymon), avec mention des numéros de volume, page(s) et colonne(s), conventionnellement cités comme suit : "FEW 16, 255a, *HRUNKJA" (dans le cas d'une référence à un endroit précis dans l'article) ou "FEW 16, 254b-255b, *HRUNKJA" (dans le cas d'une référence à un article entier). Le choix du dictionnaire en question dépend de la nature de l'unité recherchée (cf. Matthey et Thibault en préparation).

Une fois que le lexème a été localisé dans le dictionnaire, l'utilisateur qui veut connaître son étymologie doit remonter vers l'entrée de l'article où se trouve l'étymon. Le sens de consultation est donc *ascendant*. Un sens de consultation *descendant*, de l'étymon-vedette vers les données de l'article, existe également lorsque l'utilisateur recherche des informations associées à un étymon connu et, plus particulièrement, sa

[2]L'étymon est bien entendu un lexème lui aussi, mais nous réservons dorénavant le terme de lexème aux unités lexicales galloromanes dont le FEW explique l'étymologie.

[3]Par index du FEW au pluriel, nous comprenons les index situés à la fin des volumes ainsi que l'index partiel des formes (ATILF 2003). Ce dernier est, en revanche, le seul concerné lorsque nous parlons de l'index du FEW au singulier.

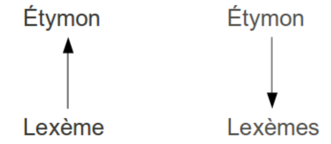


FIGURE 2.1 – Sens de consultation du FEW

descendance. Le FEW étant en effet le seul dictionnaire du domaine galloroman à rassembler sous l'étymon tous les lexèmes qui en descendent, il permet ce type particulier de recherche, qui ne peut être effectué dans les autres dictionnaires étymologiques.

Dans les volumes 21 à 23, qui regroupent les matériaux d'origine inconnue, chaque type lexical constitue, selon Büchi 1996 (77-78), un article dont l'étymon-vedette est tout simplement ellipsé (→ 2.6.1). Le lemme est donc absent en structure de surface, ce qui annule tout sens vertical de consultation. L'adresse FEW suffit pour informer l'utilisateur de l'origine inconnue du lexème. Dans cette optique, l'itinéraire de consultation de ces matériaux est plus simple que pour les matériaux étymologisés, puisqu'il consiste simplement, une fois le lexème repéré, à constater son origine inconnue. Certes, si l'on décide de considérer qu'un article, dans les volumes 21 à 23, est ce qui se trouve rangé sous un concept (analyse divergente de Büchi 1996), les deux sens de consultation ascendant et descendant valent aussi pour les matériaux non étymologisés, puisque l'utilisateur peut partir d'un concept vers les lexèmes qui sont rangés sous celui-ci, ou d'un lexème vers le concept. Il reste néanmoins que le concept n'a pas tout à fait la même valeur dans le FEW que l'étymon-vedette, puisqu'il s'agit davantage d'un moyen de classement et non d'un lemme à part entière.

Dans le FEW, l'échec de consultation est possible, par exemple lorsque le lexème n'a pas été trouvé grâce aux index. Cette absence ne signifie aucunement que le lexème est absent du dictionnaire (→ 1.1.3). L'utilisateur-consulteur doit alors deviner l'étymon et sonder l'article correspondant pour y chercher le lexème ou un lexème apparenté. Dans ce cas, les deux sens de consultation se cumulent. La consultation part d'un étymon supposé pour redescendre vers un lexème ; à partir de là, on rejoint le sens de consultation ascendant. Il est à remarquer que cette démarche heuristique tend à effacer la frontière entre consultation et lecture, puisque la recherche d'un lexème ressemble à une lecture rapide (« en diagonale ») de l'article.

Cet itinéraire double (ascendant et descendant) représente l'essentiel de ce qui est possible lorsqu'on consulte le FEW. Ce dernier ne permet pas d'autre consultation, puisque les seuls points d'entrée dans le dictionnaire sont soit un lexème (consultation ascendante), soit un étymon ou un concept (consultation descendante). Nous considérons dès lors que le lexème, l'étymon et le concept constituent les trois *unités de consultation* du FEW. Si nous laissons de côté le concept, qui possède, comme nous l'avons expliqué ci-dessus, un statut un peu particulier, nous pouvons considérer que l'étymon et le lexème représentent les extrémités de l'itinéraire vertical de consultation, que ce dernier soit ascendant (du lexème vers l'étymon) ou descendant (de l'étymon vers les lexèmes de l'article) (cf. figure 2.1).

2.3.2 Comportements de lecture

2.3.2.1 Unité de lecture

La consultation du FEW, qu'elle parte du lexème, de l'étymon ou du concept, conduit l'utilisateur dans un article du dictionnaire. L'opération suivante consiste nécessairement à lire le contenu de l'article. Cette lecture est requise même lorsque l'unique but de la consultation est de connaître l'adresse FEW d'un lexème ou d'un étymon, car il n'est pas certain qu'il s'agit bien de l'article correspondant au lexème recherché. Prenons l'exemple d'un utilisateur cherchant l'étymologie du mot français *chausse* dans le sens de « bas (du costume de la dame) ». L'index (ATILF 2003) comporte une seule entrée *chausse* au singulier, marquée comme lyonnaise, qui le mène à l'article *CASSANUS (FEW 2, 459b). Si l'utilisateur ne prend pas la peine de lire l'article, il considérera avoir trouvé le lexème et son étymon (consultation ascendante). Or, la lecture de l'article révèle que *chausse* est un mot du patois forézien (Loire) qui a comme son étymon le sens de « chêne ». Il ne s'agit donc ni du bon lexème, ni du bon étymon ! Le premier objectif de l'opération de lecture reste, dans tous les cas, de vérifier la validité de l'opération de consultation.

Le second objectif de la lecture est d'avoir accès à l'analyse effectuée par le FEW concernant le lexème, l'étymon et les relations qu'entretient le lexème (1) avec l'étymon-vedette de l'article et (2) avec les lexèmes apparentés. Dans cette optique, l'*unité de lecture*, dans le FEW, est obligatoirement l'*article* dans sa totalité.

2.3.2.2 Complémentarité des quatre champs microstructurels

L'itinéraire de lecture est, certes, globalement *descendant*, mais il se complexifie par la *lecture simultanée des notes et du commentaire*. Le lecteur doit donc nécessairement reconnaître, au sein de l'unité de lecture qu'est l'article, les quatre champs que sont l'entrée, la documentation, le commentaire et les notes (cf. Büchi 1996, 78). Parmi ces quatre champs, c'est la documentation, où se trouvent les lexèmes analysés, qui constitue le corps de l'article et l'objet principal de l'opération de lecture. Les notes, qui complètent ou commentent les données, interviennent de façon ponctuelle. Le commentaire joue quant à lui un rôle très important pour la compréhension du champ documentaire. En effet, il fournit souvent une grande part de l'analyse faite par le FEW de la famille lexicale concernée par l'article et il explicite notamment les grandes lignes de l'organisation des matériaux (cf. Büchi 1996, 154).

La complémentarité entre la documentation et le commentaire est essentielle dans l'opération de lecture. Malgré sa position en fin d'article, le commentaire se révèle souvent, en pratique, une sorte d'*introduction* et, donc, un lieu d'entrée pour la lecture de l'article. Ce rôle d'introduction joué par le commentaire ne concerne toutefois pas tous les articles. Certains d'entre eux, très courts ou à la structuration limpide, sont lisibles sans qu'un commentaire soit nécessaire. En revanche, un commentaire trop peu explicatif peut, dans le cas d'articles longs ou complexes, obliger le lecteur à lire, ou au moins parcourir, l'article en entier pour en comprendre la structure.

En fournissant une synthèse du contenu de l'article et de sa structuration, le commentaire permet par ailleurs au lecteur pressé de ne pas lire la documentation en entier, mais seulement le paragraphe ou l'ensemble de paragraphes qui l'intéresse. Dans ce

cas, le commentaire permet d'effectuer des *raccourcis* dans le parcours descendant de lecture de l'article.

2.3.2.3 Découpage hiérarchique et décodage des informations

La lecture du commentaire permet au lecteur de comprendre les grandes lignes de l'organisation des matériaux dans le champ documentaire. Or, ce dernier est souvent organisé selon plusieurs niveaux hiérarchiques, qui ne sont pas nécessairement tous explicités dans le commentaire. Par ailleurs, le format papier conduit à une recherche d'optimisation de la surface, qui se réalise à travers une condensation des informations. Le lecteur est donc amené à résoudre les nombreuses abréviations et à chercher la signification des différents sigles bibliographiques et géolinguistiques contenus dans l'article.

La lecture d'un article s'accompagne dès lors de deux opérations : 1/ le *découpage* de l'article en différents éléments hiérarchisés, qui doivent être replacés dans leur contexte ; 2/ le *décodage* proprement dit des informations, souvent abrégées. Ces opérations sont interdépendantes et suivent un itinéraire qui concerne tous les niveaux structurels du FEW, depuis la superstructure jusqu'à l'infrastructure.

Super- et macrostructure : insertion du lemme dans la structure du FEW

Avant même de lire le contenu de l'article, le lecteur doit le rattacher à ce qui le dépasse, donc le replacer dans son contexte superstructurel (organisation du dictionnaire en deux grandes parties, selon que l'origine des matériaux est connue ou non) et macrostructurel (regroupement des étymons en grandes sections linguistiques). Cette mise en contexte permet de connaître le statut étymologique des données et, pour les matériaux étymologisés, l'appartenance de l'étymon à un groupe linguistique particulier. Cette dernière information est en effet donnée par la section linguistique à laquelle appartient l'article (en pratique, par le numéro du volume ; pour une explication plus détaillée, voir Büchi 1996, 24-38). Dans certains cas, l'étymon est suivi d'une indication de langue, qui se présente sous forme abrégée et doit donc être décodée : par exemple, *andd.* signifiant *altniederdeutsch* (« ancien bas allemand »). Ce décodage s'effectue pour le lecteur non spécialiste via la consultation du *Beiheft*[4].

Microstructure : subdivision de l'unité de lecture en différents niveaux

1. *Découpage du champ documentaire*. Outre la reconnaissance des quatre champs que sont l'entrée, la documentation, le commentaire et les notes, est ici visée plus spécifiquement l'organisation du champ documentaire, qui constitue à proprement parler le corps de l'article. Le champ documentaire est en effet subdivisé en unités hiérarchisées, la plus petite étant ce que Eva Buchi appelle l'*unité minimale de traitement*, contenant le lexème et les informations qui s'y rapportent directement (Büchi 1996, 116). Ces unités minimales de traitement, ou unités lexicales, sont regroupées en divers niveaux hiérarchiques selon des critères sémantiques, morphologiques, chronologiques, géographiques et autres (cf. Büchi 1996, 105-116). Le classement diffère d'un article à

[4] L'intitulé *Beiheft*, s'il n'est pas précisé, est utilisé dans cette thèse de manière générale pour désigner le complément au FEW, et ce malgré que sa troisième et dernière édition, qui met à jour les deux autres et les remplace, soit parue en 2010 sous le titre de *Complément*.

l'autre, puisqu'il est censé rendre compte de l'histoire particulière de la famille lexicale considérée. Il est donc nécessaire de comprendre ce classement pour atteindre l'analyse offerte par le FEW.

Même si elle varie d'un article à l'autre, l'organisation d'un article en unités hiérarchisées est explicitée en partie par les marqueurs de structuration que sont les paragraphes, les marqueurs alphanumériques et la ponctuation. Le statut relativement conventionnel de ces marqueurs permet à un lecteur même non averti de saisir aisément les différents niveaux hiérarchiques de l'article.

2. *Décodage des principes organisateurs*. En revanche, les critères qui régissent cette hiérarchisation ne sont pas toujours aisés à saisir. Certains d'entre eux sont explicités par des titres de paragraphes (par exemple *Redensarten* « locutions »), d'autres par le commentaire. Dans l'extrait ci-dessous par exemple (FEW 9, 101a, PLURALIS), le commentaire final de l'article explique que le critère de regroupement des lexèmes sous b est le suffixe *-er*, qui a remplacé *-el* par analogie avec *singuler* provenant de *singularis* :

> II. 1. a. Afr. *plurel* m. «pluriel» GuernesSThomas ; adj. „qui marque la pluralité» (hap. 13. jh.). — Ablt. Afr. *pluralment* «ensemble» ChGuill, *plurellement* adv. «au pluriel» (hap. 13. jh.), *plurelment* (hap. 13. jh.), nfr. *plurelment* (ca. 1380, Aalma 9365).
> b. Afr. *plurer* m. «pluriel» (GuernesSThomas, variante ; Gillon), mfr. id. (ca. 1430), apr. *nombre pluzar* (Manosque 1293). (FEW 9, 101a, PLURALIS)

Si le commentaire ne permet pas de décoder les critères organisateurs de la microstructure, seule une lecture attentive des informations infrastructurelles (doublée de solides connaissances linguistiques) peut mener le lecteur à inférer ces critères.

Infrastructure : lecture de l'unité lexicale

L'unité lexicale se subdivise elle-même en plusieurs composants (appelés *molécules* par Büchi 1996, 116). La lecture de l'infrastructure du FEW consiste 1/ à reconnaître chacune des unités lexicales (découpage) ; 2/ à mettre en relation et à interpréter les composants qu'elles contiennent souvent sous une forme abrégée (décodage).

1. *Découpage hiérarchique*. Le découpage de l'infrastructure du FEW consiste d'abord à reconnaître où commence et où se termine chacune des unités lexicales. Ce découpage s'appuie sur la reconnaissance des molécules qui composent l'unité et, notamment, sur celle du signifiant, qui, lorsqu'il est présent en surface, constitue intuitivement pour le lecteur l'information centrale de l'unité ou, en quelque sorte, son *noyau*. Le lecteur ne place en effet pas intuitivement sur le même pied les huit molécules décrites par Büchi 1996 (à savoir l'étiquette géolinguistique, le signifiant, la catégorie grammaticale, la définition, la localisation, la datation, la référence bibliographique et les informations complémentaires). Prenons comme exemple l'extrait repris dans la figure 2.2.

Intuitivement, le lecteur distingue une structuration en trois unités, selon les signifiants *géniteur*, *geneteur* et *géniteur*, les deux premières étant regroupées sous la même étiquette géolinguistique *Fr.*. Dans la première unité, il repère deux types d'informa-

Fr. *géniteur* „celui qui engendre, père" (1137, DC; 15e s., Molin—Scarr; Hrd Cor 235), *geneteur* (1528); frm. *géniteur* „animal mâle destiné à la reproduction" (dp. Lar 1872).

FIGURE 2.2 – Lecture de l'unité lexicale (FEW 4, 102b, GENITOR)

tions : (1) tout d'abord une information de type sémantique (« celui qui engendre, père ») ; (2) ensuite diverses attestations rassemblées entre parenthèses.

Cette opération de reconnaissance des unités lexicales est complexifiée par l'ellipse en structure de surface des molécules déjà citées pour une unité précédente (règle de non répétition valable pour l'étiquette géolinguistique, le signifiant, la catégorie grammaticale et la définition ; cf. Büchi 1996, 117). Le lecteur doit dès lors prendre en compte la succession des unités dans un même paragraphe ou une même partie de paragraphe pour rétablir les composants ellipsés. Dans l'extrait ci-dessus par exemple, la définition de *geneteur* est à reprendre de l'unité précédente *géniteur*.

2. *Décodage des informations*. Une fois reconnue chaque unité lexicale, le lecteur doit l'interpréter correctement en résolvant les abréviations et, surtout, en reliant entre elles les molécules (y compris ellipsées) qui la composent. Les informations présentes à la fin de l'unité (informations complémentaires, datations, localisations ou références bibliographiques) acquièrent à ce stade une importance à ne pas négliger. Il est souvent répété que ces informations servent à fournir au lecteur un moyen de vérification, car « Wartburg tient à ce que le lecteur puisse vérifier ses matériaux » (Büchi 1996, 128). Pourtant, un certain nombre des sources bibliographiques ne sont pas accessibles au lecteur (pensons par exemple à une référence bibliographique telle que *Gl* qui, lorsqu'elle n'est suivie d'aucune mention de tome, page et colonne, renvoie non au *Glossaire des patois de la Suisse Romande*, mais au fichier inédit de ce même *Glossaire*). De même, les indications données par le FEW ne sont pas toujours assez précises pour retrouver le lexème dans une source publiée. C'est le cas par exemple de *tueyssegue*, que le FEW étiquette comme ancien occitan (*apr.*) et atteste en Provence au 15e siècle (FEW 13/2, 134a, TOXICUM), mais qui est introuvable dans les ouvrages cités comme références par le *Beiheft* pour l'ancien occitan (cf. Matthey et Thibault en préparation). En réalité, les informations d'emploi qui suivent le signifiant ne servent pas seulement à fournir un moyen de vérification. Elles jouent un rôle encore plus important, consistant à compléter l'étiquette géolinguistique en l'explicitant ou en la restreignant. Dans l'exemple de *tueyssegue*, la mention *(Provence 15. jh.)* vient restreindre la portée de l'étiquette *apr*. Par conséquent, une lecture attentive et raisonnée du FEW relie nécessairement les informations qui se trouvent avant la forme (étiquette géolinguistique) et les précisions d'emploi qui se trouvent après celle-ci.

L'interprétation des composants de l'unité lexicale est sans aucun doute l'exercice le plus complexe pour le lecteur. En effet, le décodage des niveaux supérieurs nécessitait seulement de connaître la *grammaire* du FEW, grammaire qui est normalement rapidement acquise. Le décodage de l'infrastructure nécessite en outre de maîtriser le *lexique* du FEW, qui est multiforme et abondant. De ce fait, le lecteur averti décode plus vite que le lecteur débutant, qui doit pallier son manque de connaissances lexicales par le recours au *Beiheft*, outil de déchiffrage auquel il se reporte sans cesse pour (1) résoudre les abréviations et (2) obtenir des informations supplémentaires sur un point ou une source. L'itinéraire de lecture est donc complexifié, lors du décodage de l'infrastructure, par un aller-retour constant entre le FEW proprement dit et le *Beiheft*.

2.3.3 Conclusion

L'utilisation du FEW passe donc par des itinéraires plus complexes que ceux qui valent pour un dictionnaire classique. La consultation du dictionnaire se fait soit directement par un lexème dont l'adresse FEW est connue de l'utilisateur, soit par le lemme. À partir de ce point d'entrée, la consultation suit un itinéraire descendant (du lemme à l'article) ou ascendant (du lexème au lemme). La lecture de l'article suit quant à elle un itinéraire global descendant, accompagné

- de la lecture simultanée des notes et du commentaire, ce dernier constituant en outre une sorte d'introduction à l'article et permettant des raccourcis dans le parcours de lecture ;
- d'une mise en contexte des éléments du discours, à plusieurs niveaux : insertion du lemme dans la macrostructure, découpage de la documentation en unités hiérarchisées et mise en relation, au sein de l'unité lexicale, de ce qui précède la forme (étiquette géolinguistique) avec ce qui la suit (précisions d'emploi) ;
- d'un aller-retour constant FEW-*Beiheft* pour le décodage du lexique fewien (essentiellement étiquettes géolinguistiques et sigles bibliographiques).

Ces itinéraires, qui sont suivis lors d'une lecture attentive et raisonnée du FEW, présentent des variantes selon le type d'utilisateur. Les spécialistes ont tendance à privilégier une consultation descendante parce qu'ils ont souvent une idée assez précise de ce à quoi pourrait ressembler l'étymon du lexème qu'ils cherchent. En ce qui concerne la lecture, un lecteur débutant utilise davantage le *Beiheft*, car il maîtrise moins qu'un lecteur averti le lexique du FEW. L'aller-retour FEW-*Beiheft* reste malgré tout indispensable dans une optique de vérification approfondie des données du FEW, qui est effectuée par certains utilisateurs expérimentés, ainsi que pour la critique des sources.

2.4 Itinéraires d'utilisation souhaités

La communauté scientifique réclame un FEW informatisé. Cette demande n'est pas sans rapport avec l'informatisation du TLF et pourrait donc être perçue comme un phénomène de mode, visant à profiter de nouvelles technologies pour rendre un outil plus attrayant. Il est bien connu qu'un ouvrage informatisé permet, par rapport à une version papier, de nouvelles possibilités à la fois de lecture et de consultation (→ 1.1.5).

En ce qui concerne le FEW, ces souhaits d'informatisation sont en étroite relation avec les difficultés d'utilisation de l'ouvrage : ils répondent à des besoins réels. Il est donc primordial, avant d'imaginer une informatisation, de définir exactement les besoins en question : concevoir des quantités de fonctionnalités n'aurait aucun sens si ces dernières ne résolvaient pas les problèmes rencontrés par les utilisateurs. Par ailleurs, les attentes de ces derniers par rapport à une informatisation sont intéressantes car révélatrices de la façon dont ils voudraient utiliser le FEW et, plus généralement, de la façon dont ils perçoivent l'ouvrage et ses structures.

Nous distinguons, comme dans la section précédente, les besoins relatifs à la consultation du FEW de ceux qui sont en rapport avec la lecture de l'article et le décodage des informations. Dans une troisième partie, nous abordons également les souhaits qui

concernent, non l'amélioration des stratégies d'exploitation du FEW tel qu'il est, mais l'amélioration du FEW proprement dit, c'est-à-dire la mise à jour de son contenu.

2.4.1 Souhaits en relation avec la consultation du FEW

L'analyse des comportements de consultation actuels a montré que les seules voies possibles d'entrée dans le FEW étaient les lexèmes et les lemmes (étymons essentiellement ; → 2.3.1). Cela suppose que l'utilisateur s'interroge soit sur un lexème déterminé, connu de lui (sinon sous sa forme exacte, du moins sous une forme typisée), soit sur un étymon déterminé. Or, parmi les catégories d'utilisateurs définies plus haut (→ 2.2.2), certaines sont intéressées par d'autres informations qu'un lexème ou un étymon particulier : les dialectologues par exemple pourraient trouver utile d'entrer dans le FEW par le biais de l'étiquette géolinguistique ou de la localisation.

Un type d'information tel que l'étiquette géolinguistique présente la particularité de se trouver répété à de nombreux endroits dans le dictionnaire (contrairement au lexème ou au lemme qui sont censés – sauf cas de classements multiples, qui constituent une incohérence dans le programme lexicographique – apparaître une seule et unique fois). Les lexèmes et les étymons constituent, en effet, non seulement les unités de consultation du FEW, mais également les *unités d'analyse* : les unités qu'on cherche à expliquer et autour desquelles se construit le discours fewien. Les autres types d'information – quels qu'ils soient – constituent non le sujet du discours, mais en quelque sorte le prédicat : ce qu'on dit à propos des sujets que sont les lexèmes et les lemmes. Ces informations peuvent s'appliquer à plusieurs lexèmes et donc se répéter : l'information *Malm.*, par exemple, se trouvera dans le FEW à chaque fois qu'un lexème est dit appartenir au parler de la région de Malmedy.

Une consultation du FEW qui aurait comme point d'entrée l'information *Malm.* mènerait donc, non pas à un endroit unique du dictionnaire situé dans un article précis, mais à tous les endroits du FEW où apparaît un lexème provenant de Malmedy. Ce type de recherche, qui pourrait dans notre exemple intéresser un dialectologue étudiant le parler de Malmedy, sera appelé *recherche transversale*, car il franchit les cloisons structurelles du dictionnaire pour sélectionner, sur la base d'un critère déterminé (dans notre exemple, une même étiquette géolinguistique), des unités qui appartiennent à des familles lexicales distinctes et n'ont donc pas été mises en relation directe par le discours lexicographique.

Nous avons tenté de savoir si des utilisateurs consultaient, ou désiraient consulter, le FEW de façon transversale et, le cas échéant, quels types d'information constitu(er)aient leurs voies d'entrée privilégiées dans le dictionnaire. Nous avons distingué deux types de recherches, celles portant sur un seul type d'information et celles qui combinent plusieurs critères.

2.4.1.1 Recherches transversales simples

Cherchez-vous fréquemment un ensemble de mots (galloromans) ayant une caractéristique particulière :

– une même localisation : 13 réponses positives (41 %)

– une même source : 8 réponses positives (25 %)

> – une même signification : 17 réponses positives (53 %)
> – une même langue-origine : 13 réponses positives (41 %)
> – une même caractéristique morphosyntaxique, sémiotique : 13 réponses positives (41 %)
> – autres : 11 réponses (34 %)

Les réponses au questionnaire montrent que les recherches transversales simples sont peu fréquentes dans la version papier du FEW, mais pour des raisons essentiellement pratiques. En effet, plusieurs commentaires signalent « qu'il est actuellement impossible de faire ces recherches » à moins de lire les 25 volumes du dictionnaire. Les sondés sont toutefois très enthousiastes sur l'intérêt de ce type de recherche : s'il était facilité par l'informatisation, « ce serait naturellement très utile », « cela ouvr[irait] des perspectives », « cela rendrait des services prodigieux ». Aucun des critères de recherche proposés n'est rejeté : « tout cela serait souhaitable ». Les recherches portant sur une même caractéristique sémantique apparaissent notamment comme un rêve fou qui serait « miraculeux » et « l'accomplissement du projet initial de Wartburg » (cf. FEW 1, IV-V ; Büchi et Chambon 1995, 937).

Les réponses font état de consultations du FEW qui ont été ou seraient envisagées dans ce sens. Les commentaires donnent quelques exemples pour une même localisation (« tout le lexique de l'ancien dauphinois », « tous les mots de telle localité », « tous les mots d'origine wallonne ou lorraine ») et une même origine (« tous les emprunts à l'ancien francique ») en signalant que « ce serait surtout utile pour le gaulois, le breton ou le préceltique ». Parmi les caractéristiques morphosyntaxiques sont envisagés plus particulièrement un même préfixe ou suffixe (4 réponses) et l'appartenance à un même type morphologique comme les composés verbaux (2 réponses). Sont mentionnés également l'appartenance à la catégorie des déonomastiques, « un même fait de phonétique », « toutes les graphies d'un mot », « toutes les locutions comportant [un même mot] » ou encore « tous les mots ayant un [même] type étymologique » (c'est-à-dire provenant d'étymons partageant une même caractéristique, telle qu'une même finale par exemple).

2.4.1.2 Recherches transversales complexes

> Cherchez-vous fréquemment un ensemble de mots (galloromans) :
>
> – ayant plusieurs caractéristiques particulières (« ET ») (ex. tous les mots picards datés du 13e siècle) : 13 réponses positives (41 %) un ensemble de mots (galloromans)
> – ayant l'une des caractéristiques particulières contenues dans un ensemble (« OU ») (ex. tous les mots wallons, étiquetés « liég. » ou « nam. » ou . . .) : 11 réponses positives (34 %)
> – autres : 4 réponses (13 %)

Si la recherche de lexèmes partageant une même caractéristique s'avère actuellement difficile, on s'attend au même constat pour les recherches transversales complexes, c'est-à-dire multi-critères. Elles n'ont, en effet, pas souvent été envisagées par les utilisateurs, car elles se révèlent trop compliquées et trop chronophages dans l'état actuel du FEW. Les commentaires sont toutefois positifs quant à leur intérêt : « tout cela se-

rait intéressant », « très intéressant », « ce type de recherche catégorielle serait très précieux ».

Les recherches utilisant l'opérateur « ou » apparaissent surtout d'un grand intérêt pour les mots provenant d'une même région linguistique, vu l'hétérogénéité des étiquettes géolinguistiques du FEW : « tous les mots savoyards (donc étiquetés « sav. » mais aussi avec tous les noms de commune du domaine) ». Un utilisateur signale qu'une recherche portant sur les mots wallons, par exemple, devrait automatiquement intégrer les mots pourvus d'une étiquette appartenant au domaine wallon, telles que « liég. », « nam. » etc.

Les combinaisons de plusieurs critères (opérateur « et ») envisagées dans les réponses reçues concernent quant à eux les suffixes, les sens, les dates, les catégories grammaticales, les étiquettes géolinguistiques et les étymons : « tous les mots picards formés à l'aide du suffixe XY », les « mots normands provenant du norrois », les « mots français provenant de l'occitan », « tous les mots dont les étymons se terminent par −*atem* mais non −*catem* », « tous les lexèmes francoprovençaux désignant le cresson », « tous les adjectifs occitans munis du suffixe −*enc* ».

2.4.1.3 Synthèse et mises en garde

La multiplicité des types d'information suggérés comme critères de recherche fait donc apparaître un désir fort de ce mode de consultation transversal, qui est impossible dans la version papier du FEW, mais très attendu dans l'optique de son informatisation : une réponse signale d'ailleurs que « c'est là tout l'intérêt d'un FEW informatisé ».

Toutefois, avant même de savoir si ce mode de consultation est techniquement réalisable, il faut se demander s'il est réellement souhaitable. Quelques réponses mettent en garde contre la complexité de telles recherches à critères multiples dans un FEW qui n'a pas été conçu dans cette optique. Ces mises en garde témoignent, de la part de certains utilisateurs expérimentés, d'une conscience que ce type de consultation transgresse la structuration du discours lexicographique, puisque les types d'informations recherchés ne constituent pas le sujet du discours fewien et que la syntaxe fewienne n'est pas construite autour d'eux. Un bon exemple en est que les étiquettes utilisées dans le FEW ne sont pas homogènes : elles dépendent des sources et n'ont pas du tout été normalisées dans l'optique de permettre des requêtes à partir d'elles[5], d'où l'importance de l'opérateur « ou » dans les recherches transversales complexes. Par ailleurs, quels que soient les critères de consultation utilisés, les dangers de mauvaise interprétation des résultats sont réels si les utilisateurs considèrent ce résultat comme exhaustif ou s'ils se contentent de recueillir une liste de lexèmes correspondant à leurs critères de recherche sans prendre la peine de retourner à chaque article pour lire et décoder les données suivant les itinéraires « classiques » définis au point 2.3.3.

2.4.2 Souhaits en relation avec la lecture du FEW

Nous avons vu que les itinéraires de lecture du FEW étaient complexes. Son informatisation a également été mentionnée par la communauté comme un moyen de faciliter la lecture de l'ouvrage. Ici aussi, nous avons tenté de recueillir les avis des utilisateurs.

[5] Dans l'index du FEW (ATILF 2003), les étiquettes géolinguistiques ont fait l'objet d'une normalisation.

Le questionnaire comportait à cet effet un ensemble de questions destinées à savoir quelles difficultés sont communes à la plupart des utilisateurs et quelles « aides à la lecture » seraient envisageables. Une autre partie du questionnaire cherchait à déterminer l'utilité de divers projets (qualifiés de « rêves fous » par allusion au doute émis sur leur faisabilité) tels que la traduction des commentaires allemands, la réalisation d'une cartographie automatique ou encore la mise en relation du FEW avec d'autres dictionnaires et bases de données, projets qui sont tous en rapport avec l'opération de lecture d'un article.

Les réponses à ces questions ont permis de mettre au jour six besoins spécifiques : la résolution des nombreuses abréviations du FEW, l'explicitation des étiquettes géolinguistiques, l'explicitation des sigles bibliographiques et des sources associées à une étiquette géolinguistique, la traduction des termes allemands, la mise en évidence et l'explicitation du plan de l'article et, enfin, la mise en évidence d'informations particulières.

2.4.2.1 Résolution des abréviations

> Vous paraît-il utile d'accéder directement (par « clic ») à des informations telles que :
>
> – la résolution d'une abréviation géo-historique, bibliographique, de structuration etc. (ex. liég. > liégeois, ablt. > ableitungen) : 32 réponses positives (100 %)

Trois types d'information du FEW se présentent systématiquement sous forme abrégée : les étiquettes géolinguistiques (par exemple *liég.* pour « liégeois »), les sigles bibliographiques (par exemple *Gdf* pour « Godefroy ») et les marques de structuration interne (par exemple *ablt.* pour « ableitung[en] »). L'opacité de certaines de ces abréviations constitue un problème pour les utilisateurs non spécialistes, qui doivent systématiquement consulter le *Beiheft* pour les décoder. Il est toutefois à remarquer que ce dernier ne contient pas la résolution des abréviations de structuration.

Les réponses au questionnaire sont unanimes en ce qui concerne le besoin d'un accès aisé aux résolutions des abréviations du FEW. Néanmoins, les avis divergent quant aux priorités. Dans les zones réservées aux remarques, certains utilisateurs insistent surtout sur l'importance d'une résolution des sigles bibliographiques, tandis que d'autres privilégient les abréviations géo-historiques. Ces divergences sont révélatrices du fait que la communauté des utilisateurs du FEW est constituée d'individus présentant des intérêts et des compétences variables (→ 2.2.2). Des utilisateurs expérimentés indiquent par exemple qu'ils n'ont pas besoin de la résolution de l'un de ces trois types, mais qu'ils l'estiment utile pour leurs étudiants.

2.4.2.2 Explicitation des étiquettes géolinguistiques

> Vous paraît-il utile d'accéder directement (par « clic ») à des informations telles que :
>
> – la localisation d'un sigle géolinguistique (appartenance à un groupe linguistique supérieur : ex. liég. > wallon > oïl) : 25 réponses positives (78 %)

I. Afrb. *tosche* „poison“ Gl, apr. *tueissec* m. (Provence 12.–14. jh.), *tueissigue* (Avignon 1350, R 80, 182), *tueyssegue* (Provence 15. jh.), aost. *toucho*, dauph. *tossio* (1633), Ronco *tǫ́si*, Ala *ιǫ́si*, Queyr. *tóssi*, NHengst. *tǫ́si*, Pietrap. *tṷ́asi*, Nice *tüéisəgə* „id.; ciguë“ RF 9, 264, AlpesM. *twésigə* „ciguë“ ALF 1503 p 898. Gruy. Blon. *tūtso* „aconit“, Pd'Enh. *tūtšo*, Evolène *tő̌šyo* „chèvrefeuille; fruit du chèvrefeuille“ [1]); *tōšyi* f. „chèvrefeuille“. Isère *toussio* adj. „amer, qui a un goût désagréable“ ChF. Castr. f. „prise de tabac“. – Ablt. AlpesM. *twísega* f. „poison“ [2]) ALF 1051 p 898. BAlpes *tuichié* m. „sorbier des oiseaux“ [3]) RlFl 5, 117, HUb *tuissiér*.

FIGURE 2.3 – Ensemble constituant une zone géographique homogène (FEW 13/2, 134a, TOXICUM)

La résolution des abréviations géo-historiques n'est pas toujours suffisante pour les comprendre. Par exemple, savoir que *for.* est l'abréviation de *forézien* n'est pas utile si l'utilisateur ne sait pas ce que représente le terme *forézien*. Un second problème, distinct de celui des abréviations, est donc la signification des nombreux termes techniques du discours fewien. Dans l'utilisation du FEW imprimé, ce problème est résolu par la consultation du *Beiheft*, qui explicite la plupart des étiquettes géolinguistiques (abrégées ou non) en les localisant au sein d'un ensemble plus vaste. Cette affirmation est surtout vraie pour les étiquettes de nature géographique (localité ou région) comme *Aubigné*, suivie dans le *Beiheft* de l'explicitation « (Sarthe, La Flèche, Mayet) » ou *Auge*, abréviation résolue en « Pays d'Auge oder Vallée d'Auge » et explicitée « landschaft im dep. Calvados »[6]. Dans la troisième édition du *Beiheft*, les sous-ensembles contenus dans une étiquette géographique sont également énumérés : par exemple, *LiègePr.* est explicité « province de Liège (BE) comprenant les arrondissements suivants : Verviers (y compris le canton de Malmedy), Liège, HuyL., War.) ». Les étiquettes de nature linguistique, indiquant non un lieu, mais un parler (*liég.*, *hmanc.*, etc.), sont généralement mises en contexte par le biais du lieu auquel elles se rapportent : par exemple, *hnorm.* est explicité dans la troisième édition du *Beiheft* comme le « parler de la Haute-Normandie (FR ; région correspondant aux départements de la Seine-Maritime et de l'Eure) ».

L'utilité d'un accès aisé à ces mises en contexte géographiques et/ou géolinguistiques est clairement affirmée par la majorité des sondés. Certains insistent sur le fait que ces explicitations sont surtout nécessaires pour un public non spécialiste et notamment pour les étudiants. Par ailleurs, la mise à disposition d'une carte reprenant les localisations exactes des localités subsumées par ces étiquettes est une demande esquissée à plusieurs reprises. La localisation, sur une carte, du référent d'une étiquette géolinguistique est en effet une démarche régulièrement pratiquée par les lecteurs du FEW : des cartes sont normalement fournies à cet effet en même temps que le *Beiheft* en annexe du FEW. Cette démarche sert non seulement à situer un lieu particulier, mais aussi à vérifier la proximité géographique de lexèmes mis en relation par le discours fewien mais provenant de parlers différents. Dans l'extrait de la figure 2.3 par exemple, l'ensemble des étiquettes géolinguistiques détermine une zone géographique assez homogène, située dans les domaines francoprovençal et occitan.

Dans l'optique d'aider à la réalisation de telles cartes, un projet de cartographie automatique a été imaginé au Centre du FEW. Nous avons recueilli l'avis des sondés sur ce projet :

[6]La troisième édition du *Beiheft* (2010) est encore plus explicite et indique le pays : *Auge* est restitué en « (FR, Calvados ; région) ».

> Quels projets, permettant une utilisation nouvelle du FEW, vous paraissent utiles ?
>
> – Une cartographie automatique, permettant de visualiser sur la carte du domaine galloroman les données localisées d'un article (ou d'une partie d'article sélectionnée) : 26 réponses positives (81 %), 2 réponses négatives (6 %)

L'enthousiasme exprimé dans la majorité des réponses confirme l'intérêt de la démarche cartographique pour la lecture du FEW et l'utilité d'une aide dans cette démarche. Une demande est même exprimée pour une cartographie qui permettrait de suivre l'évolution d'un phénomène à travers le temps : cette demande est révélatrice de l'utilisation du FEW dans une perspective diachronique. À l'opposé, quelques voix énergiques – et d'autorité – mettent en garde contre les dangers d'une cartographie automatique, signalant que c'est « dangereux ! » et que « [l]es utilisateurs vont faire des bêtises avec cela ». Ces inquiétudes ne sont suivies d'aucune explication, mais sont à prendre au sérieux, provenant de lecteurs avertis.

2.4.2.3 Explicitation des sources associées à une étiquette géolinguistique

> Vous paraît-il utile d'accéder directement (par « clic ») à des informations telles que :
>
> – la ou les sources(s) détaillée(s) correspondant à une abréviation géo-historique ou bibliographique : 30 réponses positives (94 %)

Dans le FEW sont associées à chaque étiquette géolinguistique une ou plusieurs sources canoniques, via le *Beiheft*. Par exemple, sous l'étiquette *AmiensE* sont citées les données du *Lexique picard des parlers est-amiénois de René Debrie* : un lexème étiqueté *AmiensE* qui ne serait accompagné dans le discours fewien d'aucune référence bibliographique proviendra nécessairement de cette source. Une lecture attentive et raisonnée du FEW ne peut donc ignorer les sources canoniques. En outre, ces informations sont particulièrement importantes pour les spécialistes qui veulent, à partir du FEW, retourner à la source des matériaux. Remarquons toutefois que l'étiquette ne permet pas à elle seule, lorsque plusieurs sources canoniques y sont associées, de savoir dans laquelle (ou lesquelles) de ces sources ont été trouvées les données : le lecteur intéressé par cette information est donc obligé de consulter les sources en question pour y vérifier la présence du lexème concerné.

Une majorité des sondés est intéressée par un accès aisé à l'explicitation des sigles bibliographiques et plus particulièrement à celle des sources canoniques correspondant à une étiquette géolinguistique. Remarquons qu'une modélisation des sources canoniques est également indispensable dans le cadre de consultations transversales dont le critère serait une référence bibliographique (→ 2.4.1).

2.4.2.4 Explicitation des sigles bibliographiques et liens externes

L'explicitation d'un sigle bibliographique revient à donner le détail complet de ses références. Cette information est également disponible pour l'utilisateur via la consultation du *Beiheft*, dans une partie distincte de celle qui donne les sources associées

aux étiquettes géolinguistiques. Ces deux parties sont donc séparées mais interdépendantes, puisqu'une source canonique peut apparaître sous la forme d'un sigle dans la partie explicative de la section concernant les étiquettes géolinguistiques : le lecteur du FEW est alors renvoyé, pour le détail des références, à la partie explicitant les sigles bibliographiques.

Certains des sondés font état de besoins qui montrent que le *Beiheft* ne répond pas à toutes les attentes, du moins dans sa deuxième édition, la plus récente de celles disponibles au moment de notre sondage. Quelques-uns demandent en effet que soit disponible une information sur la date des textes littéraires, surtout médiévaux, information qui n'est pas systématique dans le *Beiheft* et qui, de plus, dépend de l'état de la recherche scientifique. De même, pour les sources diatopiquement marquées, est demandée une information sur leur caractère dialectal ou régional, information qui n'est pas systématique dans le *Beiheft* et pose en outre pour certaines sources (par exemple le *Dictionnaire historique du parler neuchâtelois et suisse romand* de Pierrehumbert, sous *neuch.*) quelques problèmes dus à un contenu linguistique non identifié ou mixte, c'est-à-dire mêlant des données dialectales et régionales. Enfin, certains sondés demandent qu'un lien soit créé entre les sigles bibliographiques du FEW et la bibliographie du DEAF, qui fait autorité dans le domaine des sources textuelles de l'ancien français. Notons que la troisième édition du *Beiheft* répond en partie à ces attentes.

Quelques-uns des sondés émettent en outre des souhaits qui dépassent le cadre du *Beiheft* et ne sont possibles que dans la perspective d'une informatisation de l'ouvrage. Une demande est ainsi formulée afin qu'un clic sur une source informatisée donnée par le FEW comme référence à un lexème offre un accès direct à cette source, soit sur sa page d'accueil, soit directement sur la page où se trouverait le lexème concerné. Cette demande, certes un peu utopique dans l'état actuel de la numérisation des ouvrages en question, montre que les lecteurs du FEW, loin de se limiter aux références bibliographiques, utilisent ces références pour vérifier les données dans les sources. Un accès plus aisé à ces sources devient dès lors très intéressant. Sont plus particulièrement concernés les ouvrages de référence dont la consultation accompagne souvent la lecture du FEW. Nous avons tenté de savoir quels ouvrages étaient les plus souvent utilisés en parallèle au FEW :

Quels projets, permettant une utilisation nouvelle du FEW, vous paraissent utiles ?

- Des liens avec d'autres documents informatisés (ou en voie de l'être), que vous consultez parallèlement au FEW lors de vos recherches :
  - Dictionnaires : 29 réponses positives (91 %) ;
  - Bases de données : 15 réponses positives (47 %) ;
  - Autres : 6 réponses (19 %)

Les ouvrages lexicographiques sont nettement privilégiés. Les réponses citent les dictionnaires et bases de données suivants : DEAF, DMF, Frantext, TLFi, TLF-Étym pour le français général ; BDLP, DRF, DSR, DHFQ pour la lexicographie française régionale ; ensuite AND, CCFM, CNRTL, DAO/DAG, DDL, DÉRom, DOM, MED et la base des mots-fantômes. Les autres documents consultés parallèlement au FEW sont les dictionnaires dialectaux, la bibliographie du DEAF, les éditions électroniques des textes source, les textes en ancien français, les dictionnaires et bases de données analogues des autres langues romanes.

2.4.2.5 Traduction des termes techniques et des commentaires allemands

> Quels projets, permettant une utilisation nouvelle du FEW, vous paraissent utiles ?
>
> – La traduction en français des commentaires allemands : 17 réponses positives (53 %)

Le fait que la métalangue du FEW soit majoritairement l'allemand alors que l'ouvrage est utilisé en grande partie par des francophones pose évidemment un problème majeur, qui se présente surtout à deux endroits. L'allemand apparaît tout d'abord dans les marques de structuration du champ documentaire, parfois sous forme d'abréviations (*Ablt.*, *Mit suffw.*), mais aussi de façon non abrégée (*Redensarten*, *Übertragen*). C'est en outre la langue majoritairement utilisée dans le commentaire des articles.

L'objet du discours fewien, ainsi que sa nature lexicographique, expliquent que certains termes allemands reviennent fréquemment. L'intérêt d'une traduction en français de ces termes techniques est dès lors évident ; une liste constituée par des rédacteurs du FEW circule notamment dans les milieux spécialisés et est mise à disposition du public sur le site internet de l'ATILF (www.atilf.fr/few).

Le Centre du FEW possède également un fichier contenant les traductions de certains commentaires d'articles. Le projet, souvent évoqué lors de discussions informelles, de traduire en français l'intégralité des commentaires rédigés en allemand emporte l'enthousiasme de la moitié des sondés. Plusieurs réponses expliquent qu'il ne s'agit pas d'une priorité, mais que ce serait très utile pour les étudiants non germanophones. La nécessité de conserver le texte original est toutefois soulignée à plusieurs reprises, la traduction ne devant pas remplacer le texte allemand.

2.4.2.6 Mise en évidence et explicitation du plan de l'article

> Vous paraît-il utile de mettre certaines informations mieux en évidence, par exemple le plan de l'article :
>
> – 20 réponses positives (63 %)

Les commentaires en allemand méritent d'être traduits à l'attention du lecteur non germanophone dès lors qu'ils permettent de mieux comprendre la structuration des matériaux dans le corps de l'article et, par là même, d'avoir accès à l'analyse faite par le rédacteur (→ 2.3.2). La mise en évidence visuelle des marqueurs de cette structuration participe du même objectif et concerne tous les lecteurs du FEW, y compris germanophones. La lisiblité de ces marqueurs paraît toutefois ne pas présenter de problème particulier. Plusieurs sondés indiquent que les marqueurs structurels sont assez visibles tels qu'ils se présentent dans le FEW, malgré sa typographie dense. En revanche, le besoin d'une explicitation des critères de structuration, qui varient fortement d'un article à l'autre, est clairement affirmée. Ce besoin a notamment été pris en compte, sur l'exemple du LEI, dans la refonte des articles de la tranche alphabétique B- : les articles BILANX et BASTUM, par exemple, débutent par un plan de l'article dans lequel les critères de structuration sont clairement donnés (www.atilf.fr/few). En ce qui concerne

les articles imprimés du FEW, cette demande implique une mise en relation des marqueurs apparaissant dans la documentation avec leur explicitation dans le commentaire de l'article (→ 2.3.2).

2.4.2.7 Mise en évidence d'informations spécifiques

> Vous paraît-il utile de mettre certaines informations mieux en évidence, par exemple les étymons cachés :
>
> – 26 réponses positives (81 %)
> – autres : 5 réponses (16 %)

La présentation condensée du FEW papier le rend peu lisible. Certaines informations, particulièrement importantes pour tout ou partie des utilisateurs, sont noyées dans la densité de la mise en page. C'est le cas notamment des sous-lemmes, dont la reconnaissance par le lecteur est capitale : elle permet en effet de lever la contradiction entre présentation lexicographique et analyse lexicologique qui caractérise un nombre non négligeable d'étymologies du FEW (cf. Büchi 1996, 52). Büchi 1996, 405-564 relève une liste impressionnante de ces étymons cachés, dont la mise en évidence au sein du texte fewien serait très utile d'après 80 % des sondés. Une des réponses signale qu'un besoin encore plus crucial que leur mise en évidence visuelle est la possibilité de les rechercher grâce à leur indexation.

Les sondés émettent en outre le souhait d'une mise en évidence visuelle d'autres informations telles que les suffixes, les renvois entre articles, les données romanes non galloromanes, ainsi que les données non romanes. Ces deux dernières demandes sont révélatrices de l'utilisation du FEW hors du domaine galloroman (→ 2.2).

2.4.2.8 Synthèse et mises en garde

Les six besoins décrits ci-dessus peuvent être rassemblés en trois catégories. La première concerne le besoin de résolution et d'explicitation des *termes techniques* du FEW que sont les étiquettes géolinguistiques, les sigles bibliographiques et les marqueurs de structuration de la partie documentaire. Ce besoin confirme la difficulté éprouvée par certaines catégories d'utilisateurs à lire le FEW en raison de l'abondance de son lexique technique. Nous pouvons également inclure dans cette catégorie l'explicitation des sources canoniques associées à une étiquette géolinguistique, besoin qui concerne tous les utilisateurs sans distinction. Cette catégorie est partiellement (marqueurs de structuration exclus) couverte par le *Beiheft*. La traduction en français de son contenu, effectuée à l'occasion de la troisième édition, a pour but de le rendre plus accessible. La traduction des marqueurs de structuration participe à une deuxième catégorie de besoins, qui consiste à expliciter la *structuration* du discours lexicographique. Ce besoin est exprimé par la demande d'un plan de l'article et par la traduction des commentaires allemands. L'explicitation du plan de l'article requiert en effet une mise en relation des marqueurs alphanumériques du champ documentaire avec leur signification donnée dans le commentaire. Enfin, une troisième catégorie consiste à ouvrir le FEW *vers l'extérieur* en facilitant la consultation d'outils complémentaires.

Certains besoins sont plus spécifiques à certaines catégories d'utilisateurs. La résolution des abréviations et l'explicitation des étiquettes géolinguistiques concernent davantage les lecteurs débutants et notamment les étudiants. La traduction des termes techniques allemands concerne essentiellement les lecteurs non germanophones. En revanche, d'autres besoins, comme l'explicitation des sources associées à une étiquette géolinguistique, sont communs à l'ensemble de la communauté, spécialistes y compris.

La plupart de ces besoins sont en rapport direct avec la demande d'informatisation du FEW. Ils se résolvent en effet par des *mises en relation* (avec le commentaire, avec le *Beiheft*, avec des outils externes) qui, certes, sont possibles dans la version imprimée du FEW, mais seraient grandement facilitées par une informatisation de son contenu. La création de liens directs entre les termes techniques du FEW et leur explicitation dans le *Beiheft* n'est possible qu'à partir du moment où l'on informatise les deux ouvrages. Il en va de même pour la consultation d'outils externes au FEW, qui serait encouragée si tous ces ouvrages étaient disponibles sous forme électronique et interrogeables. Enfin, le report sur une carte géographique des matériaux fournis par un article ou une partie d'article serait également plus rapide dans le cadre d'une informatisation du FEW.

Certains de ces souhaits provoquent toutefois des mises en garde. Tout d'abord, quelques commentaires insistent sur la nécessité de veiller à l'ergonomie, en demandant que toutes les explicitations fournies n'alourdissent pas la présentation visuelle de l'article. Les inquiétudes les plus sérieuses concernent ensuite l'idée d'une cartographie automatique. Les problèmes que poserait un tel projet sont notamment à mettre en rapport avec l'hétérogénéité des étiquettes géolinguistiques du FEW, qui sont de nature diverse (géographique ou linguistique) et, surtout, qui peuvent référer à des sources diverses, donc à des réalités linguistiques légèrement différentes en fonction de la source concernée (→ 2.4.2). Ces caractéristiques du discours fewien rendent particulièrement critiquable la demande d'une cartographie diachronique. Les mises en garde sont totalement justifiées et rappellent celles émises à propos des consultations transversales (→ 2.4.1) : le danger de mauvaise interprétation des données est réel si les utilisateurs se contentent du résultat fourni par une cartographie automatique sans retourner au texte fewien pour analyser les données et les mettre en contexte, suivant les itinéraires de lecture définis plus haut (→ 2.3.2). Par ailleurs, une mise à jour du *Beiheft* était demandée, avec raison, avant de mettre en place une telle application (rappelons que le questionnaire a été distribué avant la parution de sa troisième édition). Cette mise à jour du *Beiheft* était également requise pour les demandes d'explicitation des termes techniques du FEW.

2.4.3 Souhaits en relation avec la mise à jour du FEW

Jusqu'à présent, nous avons examiné les difficultés qu'éprouvaient les utilisateurs lors de la consultation et de la lecture du contenu du FEW. Par « contenu du FEW », nous entendons les articles des volumes 1 à 25. Or, ce contenu est régulièrement revu et critiqué par les spécialistes et se trouve dès lors en partie obsolète. Par exemple, de nombreux matériaux considérés par le FEW comme d'origine inconnue ou incertaine ont depuis fait l'objet d'étymologisations. Pour en prendre connaissance, l'utilisateur du FEW doit consulter un grand nombre d'études annexes, qui se présentent sous des formats divers (livres, articles, dictionnaires, comptes rendus etc.) et ne sont pas aisément accessibles en dehors des bibliothèques spécialisées. Encore faut-il que l'utilisateur soit au courant de l'existence de ces études.

Consciente de ce problème, la communauté a émis le souhait d'une mise à jour du FEW qui intégrerait toutes les critiques apportées à l'ouvrage depuis sa parution. Il ne s'agit donc pas seulement de donner aux utilisateurs un accès plus aisé aux études annexes (ce qui serait une mise en relation comparable à celles décrites au point 2.4.2), mais d'améliorer le contenu même du FEW en y intégrant le résultat de ces études. Une partie du questionnaire était destinée à recueillir les souhaits des utilisateurs à ce sujet. En effet, il est nécessaire, d'une part de connaître la nature des informations dont ils voudraient une mise à jour, d'autre part de déterminer précisément la façon dont ils envisagent l'intégration de ces mises à jour au sein du discours fewien.

2.4.3.1 Nature des ajouts et corrections

Si vous pouviez effectuer des apports au FEW, quels seraient-ils ?

– ajout d'une information portant sur un lexème : 26 réponses positives (81 %)
– ajout d'un lexème : 17 réponses positives (53 %)
– autres : 7 réponses (22 %)

Améliorer le contenu du FEW consiste tout d'abord à y apporter des ajouts. Les réponses au questionnaire insistent sur le fait que le FEW n'est jamais achevé et qu'un apport continuel de nouvelles données est inévitable. Des spécialistes disposent d'informations actualisées sur tel ou tel lexème étudié dans le cadre de leurs recherches et trouveraient intéressant d'en faire profiter la communauté. Selon les sondés, ces informations portent notamment sur des datations nouvelles, sur des localisations nouvelles ou sur des données non galloromanes. Un apport particulier proposé dans les réponses reçues consiste à spécifier le caractère dialectal ou régional d'un lexème (désambiguisation des étiquettes géolinguistiques).

L'ajout de nouveaux lexèmes est notamment envisagé pour certains domaines techniques mal représentés dans le FEW. Enfin, dans la rubrique « autres » sont mentionnés la parémiologie, l'ajout de nouveaux étymons (donc de nouveaux articles), la suppression de lexèmes et l'ajout de précisions dans les définitions[7].

Si vous pouviez effectuer des corrections dans le FEW, quelles seraient-elles ?

– correction de la datation d'un lexème particulier (antédatation, rétrodatation, postdatation) : 24 réponses positives (75 %)
– correction de l'étymologie d'un lexème ou d'un ensemble de lexèmes : 18 réponses positives (56 %)
– réécriture d'articles : 11 réponses positives (34 %)
– autres : 3 réponses (9 %)

En ce qui concerne les corrections susceptibles d'être apportées au FEW, les sondés mettent en évidence leur nombre important et regrettent qu'elles ne soient pas mises actuellement à la disposition de la communauté. Les corrections de datations sont particulièrement concernées. Les sondés mentionnent aussi les corrections de sens. Les

[7] L'évaluation critique des sources du FEW est également évoquée dans cette rubrique par un des sondés, mais il ne s'agit pas d'un ajout au sens strict où nous l'envisageons ici.

corrections d'étymologies aboutissent à une conséquence intéressante dans le FEW, puisqu'elles peuvent conduire à un transfert d'un lexème vers un autre endroit (autre paragraphe du même article, autre article de la même partie, autre partie) du dictionnaire : les réponses font notamment allusion à la réintroduction des matériaux d'origine inconnue, suite à leur étymologisation, dans la partie étymologique du FEW.

2.4.3.2 Modalités d'intégration

En ce qui concerne la façon dont les mises à jour seraient intégrables dans le FEW, le questionnaire posait trois questions qui concernent respectivement le lieu, l'auteur et l'intervalle de publication.

> Où voudriez-vous que se fasse la publication des corrections proposées ?
>
> – directement dans le FEW : 17 réponses positives (53 %)
> – en dehors du FEW, mais regroupées en un seul endroit : 14 réponses positives (44 %)

Les réponses à la question du lieu de publication sont partagées. Davantage de sondés préfèrent une intégration des mises à jour au sein du discours fewien. Deux sondés envisagent les deux solutions simultanément, selon la nature de l'ajout : les corrections ponctuelles seraient à effectuer dans le texte du FEW et les corrections plus importantes ailleurs. Parmi les réponses qui préfèrent une insertion directe dans le texte, un consensus se dégage toutefois sur la nécessité de bien distinguer les modifications en tant que telles, afin de ne pas les confondre avec le texte original. En cela, les tenants de l'insertion directe rejoignent les préoccupations des sondés qui jugent préférable une publication hors FEW.

> Par qui ?
>
> – Par celui qui propose la correction : 5 réponses positives (16 %), dont 4 (13%) en duo avec la suivante
> – Par soumission à un expert (rédaction du FEW) qui validerait et intégrerait la correction : 31 réponses positives (97 %)

Concernant la question du responsable de la mise à jour, quatre réponses seulement demandent que la correction soit publiée par son auteur, et trois d'entre elles ont coché également la nécessité de recourir à une validation par un expert. La soumission à un expert apparaît donc à une quasi-unanimité comme « essentielle » et « absolument nécessaire ». La validation par un expert constitue certainement un garde-fou assez fiable pour éviter le risque d'intégration de mauvaises corrections ou de corrections placées au mauvais endroit.

> Quand ?
>
> – Ponctuellement, au fur et à mesure des propositions : 24 réponses positives (75 %), dont 2 (6 %) en duo avec la suivante
> – À intervalles réguliers (lesquels ?) : 8 réponses positives (25 %), dont 2 (6 %) en duo avec la précédente

En ce qui concerne l'intervalle de publication, une intégration ponctuelle des modifications emporte les préférences : « c'est là tout l'intérêt d'une version informatisée ». Les autres réponses envisagent une publication tous les ans ou tous les deux ans. Deux sondés envisagent les deux solutions, en fonction de la nature des corrections : les corrections de détail, concernant par exemple les datations, pourraient être intégrées au fur et à mesure, tandis que les corrections plus importantes feraient l'objet de plublications plus espacées.

Il est à noter que la plupart des réponses envisageant une intégration des corrections à l'intérieur du dictionnaire sont également favorables à une publication ponctuelle.

2.4.3.3 Synthèse et mises en garde

La question de la mise à jour du FEW intéresse plus particulièrement les utilisateurs avertis et, parmi eux, les chercheurs en linguistique et en philologie, ainsi que les rédacteurs du dictionnaire eux-mêmes : en résumé, les spécialistes qui, dans leur recherche, sont amenés à corriger le FEW ou à y apporter de nouvelles données. Cette question intéresse aussi les utilisateurs non avertis dans le sens où actuellement, la masse de corrections et d'ajouts au FEW se trouve consignée en dehors de l'ouvrage, dans une grande variété d'études annexes dont ils ne soupçonnent parfois pas même l'existence et qui ne sont pas facilement disponibles en dehors des bibliothèques spécialisées.

La nature des ajouts et corrections témoigne des divers domaines d'étude des spécialistes : la philologie (plus particulièrement ajout et correction de datations provenant de l'étude des sources littéraires), la dialectologie (plus particulièrement ajout et correction de localisations et spécification du caractère dialectal d'un lexème), la linguistique romane (ajout et correction de données romanes autres que galloromanes), la sémantique (ajout de nouveaux sens, ajout de précisions dans les définitions) etc.

Si la nécessité d'un accès plus aisé à ces ajouts et corrections est indéniable, les avis recueillis font montre d'une grande prudence quant aux modalités d'intégration dans le FEW de ces nouvelles données. Un commentaire signale que tous les ajouts ne sont peut-être pas souhaitables car « le FEW est tellement cohérent dans ses qualités et ses défauts qu'il ne faut pas l'entamer ». Il s'agit là d'une mise en garde sérieuse, qui tend à refuser une mise à jour complètement intégrée dans le discours fewien : les ajouts seraient en effet « à placer 'hors texte' pour ne pas créer la pagaille » ou à distinguer du texte original de façon visible. Un système d'historique des modifications est également demandé. La perspective d'un FEW évolutif entraîne la peur, non seulement de voir disparaître le texte original, mais surtout de briser la cohérence du FEW et la rigueur scientifique de son discours. Plusieurs des avis reçus font ainsi apparaître une vision du FEW comme objet historique, formant un tout cohérent, qu'il serait dangereux de modifier. Un commentaire signale le « risque de voir un FEW mal corrigé se substituer à un FEW incomplet mais de plus grande rigueur scientifique ».

2.4.4 Conclusion

Les souhaits des utilisateurs sont révélateurs de la façon dont ils utilisent, ou voudraient utiliser, le FEW. L'enquête avait pour but de savoir quels itinéraires de consultation et de lecture du FEW devraient être facilités et de voir si d'autres itinéraires étaient souhaités.

En ce qui concerne la consultation du FEW, les souhaits des utilisateurs induisent des itinéraires totalement nouveaux par rapport aux itinéraires traditionnels. En effet, les points d'entrée dans le dictionnaire ne sont plus réduits aux seuls lexèmes et lemmes (unités de consultation classiques et également unités d'analyse), mais s'étendent à tout type d'information. Ce type de consultation transversale mène à plusieurs endroits en même temps dans le dictionnaire. Ces nouveaux itinéraires, impossibles dans la version imprimée du FEW, sont très attendus par la communauté. Ils provoquent toutefois la crainte, venant de spécialistes conscients des défauts de l'ouvrage, que les utilisateurs non avertis se contentent de recueillir une liste de lexèmes correspondant à leur recherche et se servent de ce résultat sans l'analyser par un retour aux articles dont ils proviennent.

Il nous semble que les effets indésirables induits par une consultation transversale peuvent être réduits si l'on opère une distinction rigoureuse et nette entre les itinéraires de consultation, c'est-à-dire de voie d'entrée dans le dictionnaire, et les itinéraires de lecture. Les itinéraires de consultation peuvent utilement être améliorés de façon à répondre aux besoins des utilisateurs. L'élargissement des possibilités en termes de consultation ne devrait toutefois pas impliquer des changements dans la façon dont les utilisateurs lisent – ou devraient lire – un article du FEW.

La pertinence de nouveaux comportements de lecture n'est en effet pas affirmée. Les souhaits les plus partagés concernent surtout l'optimisation des itinéraires actuels. Les premiers concernés sont la lecture simultanée du commentaire et l'aller-retour constant entre le FEW et le *Beiheft* : l'intégration dans le corps de l'article, via les termes techniques et les marqueurs de structuration, des explicitations fournies par le *Beiheft* et par le commentaire rendraient ces parcours de lecture moins fastidieux. La mise à disposition d'un plan, même non explicatif, de l'article et la traduction des commentaires allemands participe également à l'optimisation des itinéraires traditionnels. Enfin, la mise en relation du FEW avec les outils externes que sont d'autres dictionnaires ou bases de données informatisés est également une démarche pratiquée par les utilisateurs de la version papier. Ce besoin est révélateur d'un comportement de lecture critique du FEW, caractéristique des lecteurs avertis, qui vérifient et continuent l'analyse du FEW en retournant à la source. De ce fait, la mise en relation des sigles bibliographiques du FEW avec les sources disponibles sous forme électronique présenterait l'avantage d'inciter les lecteurs non avertis à adopter une même démarche critique, puisqu'elle serait encouragée par l'informatisation. La constitution de cartes automatiques ne dépasse pas non plus les itinéraires traditionnels, puisque la localisation, sur une carte, d'une étiquette géolinguistique est une démarche (qui devrait être) pratiquée par les utilisateurs du FEW : une carte est normalement fournie à cet effet en annexe du FEW, avec le *Beiheft*.

Le souhait d'informatisation du FEW n'a donc pas pour objectif de modifier les itinéraires de lecture classiques de la version imprimée. Il consiste essentiellement à *faciliter la mise en relation de données* qui, dans le discours lexicographique, ne sont pas situées côte à côte. Ce faisant, il mène à des *parcours de lecture hypertextuels* qui n'étaient pas identifiés comme tels dans l'analyse des comportements d'utilisation du FEW papier, mais qui étaient sous-jacents. La complexité des itinéraires de lecture traditionnels est, en effet, due à ces mises en relation et mises en contexte obligatoires pour s'approprier l'analyse approfondie des données fournie par l'ouvrage (→ 2.3.3). L'informatisation souhaitée par les utilisateurs aurait pour conséquence principale et positive d'encourager les lecteurs à effectuer ces mises en relation nécessaires.

Ce comportement a son pendant négatif, qui consisterait, en naviguant ainsi dans un FEW hypertexte, à emprunter des raccourcis qui ne rendent pas compte de la contextualisation des données. Dans ce cas, l'hypertextualité, au lieu d'optimiser les itinéraires, nuirait à la lecture du FEW. Le même danger menace l'utilisation de cartographies automatiques. En effet, une lecture raisonnée du FEW relie toujours l'étiquette géolinguistique aux précisions d'emploi qui suivent la forme (→ 2.3.2). L'élaboration de cartes automatiques, faites à partir des seules étiquettes géolinguistiques, ne contient pas cette analyse. Il s'agit donc d'un raccourci qui peut mener à de mauvaises interprétations des données du FEW si le lecteur n'effectue pas une analyse critique de la carte en retournant aux données du FEW et, plus précisément, aux informations bibliographiques et autres citées après chaque forme. Or, la création de telles cartes risque de faire oublier, même aux lecteurs avertis, la nécessité de toujours effectuer une contextualisation des données fournies par le FEW. L'informatisation est une arme à double tranchant, qui peut, ou bien encourager les « bons » comportements de lecture, ou bien les faire oublier en les court-circuitant ; tout dépend de la façon dont on s'en sert.

2.5 Deux visions du FEW

2.5.1 Des divergences apparentes

L'analyse des souhaits des utilisateurs a mis en lumière des consensus, mais aussi des divergences : des besoins manifestes et largement partagés se heurtent à des mises en garde sérieuses. Les dissensions apparaissent autant en ce qui concerne la consultation du FEW (consultations transversales) que sa lecture (cartographies automatiques) et sa mise à jour (intégration des ajouts dans le texte fewien). Ces trois projets soulèvent en effet critiques et inquiétudes chez les uns, enthousiasme débordant chez les autres. L'explication de ces opinions contradictoires n'est pas à chercher du côté des catégories d'utilisateurs du FEW, puisqu'elles apparaissent même chez des lecteurs avertis de la même sous-discipline.

Il nous semble que l'explication se trouve dans le FEW lui-même. La consultation transversale du FEW pose problème pour deux raisons, d'abord parce que le FEW n'est pas du tout conçu dans cette optique, ensuite parce que le résultat de cette consultation n'est pas fiable si l'utilisateur ne prend pas la peine de le vérifier en lisant les articles d'où proviennent les lexèmes. Le projet de cartographie automatique suscite des réserves pour les mêmes raisons : l'hétérogénéité des étiquettes géolinguistiques rend en effet l'opération périlleuse, et le danger existe que le lecteur interprète la carte sans effectuer la relation nécessaire entre l'étiquette géolinguistique et les conditions d'emploi qui la précisent ou la restreignent (→ 2.3.2). Dans le cas d'une mise à jour intégrant les ajouts directement dans le texte original, le danger est de briser la cohérence syntaxique du FEW garante des relations précitées.

Les mises en garde insistent donc sur l'importance de la syntaxe fewienne comme élément essentiel pour s'approprier l'analyse complète des données effectuée par le rédacteur. Elles sont justifiées par la vision de l'article du FEW comme un discours construit et structuré, dans lequel chaque information est à mettre en relation avec celles qui l'entourent. Rappelons que les articles du FEW, classent et hiérarchisent les données différemment de façon à retracer l'histoire particulière de chaque famille lexicale, ce qui fait de chaque article une monographie à part entière :

> L'ouvrage se présente, en fait, comme un ensemble structuré de monographies, dont la forme lexicographique n'est qu'un auxiliaire au service de la « visée globalisante » (Swiggers 1990 : 347) de Wartburg, qui l'anime et la domine. (Büchi et Chambon 1995, 952)

Cette particularité explique que les articles du FEW se lisent davantage qu'ils ne se consultent, selon des itinéraires qui nécessitent un découpage de l'article en unités hiérarchisées, leur mise en relation et leur compréhension par la lecture simultanée du commentaire (→ 2.3.2). Les mises en garde visent donc essentiellement à protéger ce que nous appellerons la *dimension monographique* (ou *dimension M*) du FEW, qui n'est accessible que par l'opération de lecture. Les demandes, partagées par les utilisateurs, d'un plan de l'article et d'une traduction du commentaire sont tout à fait liées à cette dimension monographique du FEW : elles visent à atteindre plus aisément le classement des données et l'analyse qui en découle.

Comment s'explique, alors, l'enthousiasme des sondés, dont certains sont pourtant des lecteurs avertis conscients de cette dimension monographique du FEW, pour des projets qui ne préservent pas la cohérence du discours fewien ? Leur enthousiasme est à expliquer par une autre vision de l'ouvrage, que nous appellerons la *dimension thesaurus* (ou *dimension T*) du FEW. Dans cette vision du FEW comme un thesaurus, les utilisateurs sont intéressés par la masse de données qui s'y trouve et par les informations qui sont associées à chaque lexème. C'est le lexème, et non plus l'article, qui constitue leur centre d'intérêt. Le FEW est en effet le seul dictionnaire où se trouvent rassemblés tous les lexèmes des langues et dialectes du domaine galloroman, ce qui en fait un ouvrage des plus précieux dans de nombreuses sous-disciplines linguistiques. Il a même été considéré comme un trésor des patois, ce qu'il n'est pas (cf. Büchi 1996, 90).

Cette dimension thesaurus est concernée par les demandes de consultation transversale, qui permettraient d'accéder directement à un groupe de lexèmes partageant un point commun malgré leur dispersion lexicographique due à leur appartenance à des familles lexicales différentes ; par les besoins de mise en relation avec le *Beiheft* et avec d'autres dictionnaires, qui permettraient d'accéder directement à l'intégralité des références bibliographiques associées à un lexème ; ainsi que par les demandes d'intégration dans le FEW des ajouts et corrections apportées ailleurs. Enfin, c'est cette dimension qui explique que le FEW se consulte généralement de façon ascendante à partir des lexèmes, alors que les entrées sont des étymons.

2.5.2 Complémentarité des deux dimensions

La double vision du FEW qu'en ont ses utilisateurs n'est finalement que le reflet de la façon dont l'auteur a conçu son œuvre. Ce que nous appelons la dimension monographique du FEW est en effet le résultat de l'esprit de synthèse du maître (cf. Büchi et Chambon 1995, 945) et, en particulier, du choix d'une étymologie intégrante. La dimension thesaurus est quant à elle mise en évidence par le sous-titre du FEW (*thesaurus galloromanicus*) et assumée par l'auteur : « De l'aveu même de Wartburg, la "valeur de l'FEW, si valeur il y a, consiste surtout dans le fait d'être un Thesaurus Galloromanicus" (1959, 208). » (Büchi 1996, 130)

Cette dimension thesaurus a été critiquée (Malkiel 1976 ; cf. Büchi 1996, 130), mais représente la base sur laquelle Wartburg fonde ses étymologies. Elle est à considérer

comme un moyen et non comme une fin en soi : « [...] c'est bien à partir de ses longs listings en partie répétitifs que le FEW bâtit ses étymologies, tout comme sa conception de l'étymologie, qui est plus englobante que celle de Malkiel. » (Büchi 1996, 131)

La dimension T explique la visée totalisante et exhaustive des données et la séparation entre le commentaire et le champ documentaire, dans lequel les matériaux sont censés être donnés de façon objective (Büchi 1996, 133-135). L'opposition entre les deux dimensions recouvre ainsi partiellement l'opposition entre le champ documentaire (recueillant les matériaux) et le champ du commentaire (expliquant l'organisation de l'article) ou, dans la documentation, l'opposition entre l'infrastructure (contenant les lexèmes et les informations directement associées) et la microstructure (hiérarchisant les unités). La dimension thesaurus se focalise sur les lexèmes vus comme individus, tandis que la dimension monographique les organise de manière à expliquer leurs relations. Par ailleurs, la dimension M ne concerne pas seulement l'organisation des matériaux à l'intérieur de chaque article (microstructure) : on peut l'élargir à la macrostructure du FEW, qui consiste également en une analyse des données censée rendre compte de leur étymologie (groupement des matériaux par section linguistique de leur étymon).

Les termes de ces oppositions montrent que les deux dimensions ne sont pas inconciliables, mais étroitement imbriquées et complémentaires dans la construction du discours fewien. Nous pensons dès lors que la double vision qu'ont du FEW ses utilisateurs est légitime. Le danger survient lorsqu'ils privilégient une des deux dimensions, en particulier la dimension T, au détriment de la seconde. Par exemple, la mise à jour du FEW par intégration des ajouts dans le texte n'est pas critiquable en soi : les inquiétudes proviennent dès lors que cette intégration porte atteinte à la cohérence syntaxique du discours monographique. De même, les consultations transversales sont utiles pour accéder rapidement à un ensemble de lexèmes éparpillés dans le dictionnaire : un problème se pose uniquement si ces lexèmes ne sont pas ensuite replacés dans leur contexte au sein de chaque article. Si la consultation gagne à être envisagée dans une dimension thesaurus où chaque lexème est individualisé et accessible séparément, la lecture, quant à elle, ne peut s'effectuer que dans une dimension monographique, c'est-à-dire dans une mise en relation et une contextualisation des données.

2.6 L'implicite fewien

Un problème majeur qui se pose régulièrement à l'utilisateur du FEW est la difficulté d'accès à certaines données. Ne sont pas concernées ici les difficultés de consultation (telle que la recherche d'un lexème non indexé), mais les difficultés survenant lors de la lecture d'un article (→ 1.1.1). Ce problème d'inaccessibilité des données, généralement attribué aux structures complexes de l'ouvrage ou à sa présentation très condensée, est résumé sous le terme d'*implicite*. Nous proposons d'analyser cette question afin de déterminer les causes du problème et les moyens qui sont à la disposition de l'utilisateur pour le résoudre.

2.6.1 Relevé de l'implicite

Büchi 1996 parle à plusieurs reprises d'informations « implicites » ou « sous-entendues ». Parmi les informations catégorisées d'implicites, se trouvent

– l'origine inconnue d'un lexème, information qui découle implicitement de la position du lexème dans les volumes 21 à 23 ;

– les molécules de datation et de référence bibliographique implicitement contenues dans certaines étiquettes géolinguistiques (Büchi 1996, 126) ;

– les précisions par rapport à l'étiquette géolinguistique implicitement données par la molécule de référence bibliographique (Büchi 1996, 128) ;

– les critères de regroupement des lexèmes dans le champ documentaire (Büchi 1996, 105-116 ; 108) ;

– les corrections d'étymologie implicites car situées à d'autres endroits du dictionnaire (Büchi 1996, 19) ;

– le rédacteur d'un article, qui est implicitement von Wartburg en cas d'absence de signature explicite[8].

Pour d'autres informations, Éva Buchi ne parle pas d'implicite, mais d'ellipse. Ce terme renvoie à l'analyse qu'elle fait d'une structure du FEW modulaire et subdivisée en deux niveaux : une structure de surface et une structure profonde. Il y a ellipse lorsqu'un élément structurel (ou module) est absent en surface, mais présent en structure profonde (cf. Büchi 1996, 77-78 ; 98). Les informations suivantes sont ellipsées dans certains contextes :

– le champ de l'entrée, dans les volumes 21 à 23 (Büchi 1996, 79) ;

– la langue dont relève l'étymon, quand il s'agit de la langue par défaut de la section linguistique où se trouve l'article (Büchi 1996, 80) ;

– la glose de l'étymon, quand elle aurait la même forme que celle de l'étymon (Büchi 1996, 86) ;

– les molécules « obligatoires » (étiquette géolinguistique, forme, catégorie grammaticale et définition), quand elles constitueraient une redondance par rapport à l'unité précédente (Büchi 1996, 117 parle dans ce cas de *sous-entendus*) ;

– le champ du commentaire, dans une grande partie des articles des volumes 21 à 23 (Büchi 1996, 137).

Ces informations ellipsées sont à inclure dans l'ensemble des informations implicites du FEW. Par exemple, lorsque la langue de l'étymon est ellipsée, elle est implicitement la même que celle par défaut de la section.

2.6.2 Analyse structurelle

Dans tous les cas cités ci-dessus, l'implicite apparaît lorsqu'une information se trouve placée *à un autre endroit* du dictionnaire que là où elle est attendue. Par exemple, la langue de l'étymon est implicite lorsqu'elle n'est pas donnée juste après l'étymon (situation attendue), mais est fournie indirectement par la section linguistique (ou numéro de volume) dans laquelle se trouve l'article.

[8]Dans la section slave du volume 20, le rédacteur par défaut est Jänicke.

Cette distance entre l'endroit où serait attendue une information et celui où elle se trouve réellement est souvent provoquée dans le FEW par sa dimension monographique, qui, en organisant les données, déplace à un niveau supérieur des informations qui concernent plusieurs éléments et qui, sinon, seraient répétées autant de fois qu'il y a d'éléments concernés. Cette conséquence de la dimension monographique du FEW est évidemment problématique dans une utilisation thesaurus, où l'on voudrait que toutes les informations concernant un lexème donné y soient directement associées.

Ceci étant dit, cette « mise à distance » n'est pas réservée dans le FEW aux informations citées ci-dessus. Les explicitations des sigles situées dans le *Beiheft* partagent aussi cette propriété. De même, les corrections et ajouts externes posent des problèmes d'accessibilité parce qu'ils sont situés à un autre endroit (hors du FEW) que là où ils devraient se trouver s'ils étaient intégrés dans le dictionnaire.

Prenant tous ces cas en compte et élargissant le terme d'implicite à toute information non située là où on l'attendrait, nous pouvons établir une typologie des données implicites selon l'endroit où elles se trouvent. Nous distinguons ainsi, de la plus courte à la plus grande distance entre l'endroit attendu et l'endroit effectif :

1. les informations qui se trouvent dans la structure de surface du FEW ;
2. les informations qui se trouvent dans sa structure profonde ;
3. les informations qui se trouvent dans le *Beiheft* (résolution des abréviations, détails des sources bibliographiques, localisation d'une étiquette géolinguistique, sources correspondant à une étiquette géolinguistique, datation d'une source etc.) ou en fin de volume (corrections) ;
4. les informations qui se trouvent en dehors du FEW.

La distance lexicographique n'est pas nécessairement corrélée à la facilité de résolution par l'utilisateur. Une information située en structure profonde est ainsi plus difficile à atteindre par un lecteur débutant qu'une information située dans le *Beiheft*. En effet, les données du *Beiheft* sont accessibles directement par une mise en relation terme à terme, tandis qu'une information en structure profonde nécessite une connaissance approfondie de la façon dont est construit le discours fewien. De même, les informations situées hors FEW peuvent être très accessibles dès lors que l'utilisateur a ces documents à sa disposition.

2.6.3 Analyse du point de vue du décodage

Si l'on se place du côté de l'utilisateur qui doit décoder l'implicite, on peut distinguer l'implicite qui est inférable de l'implicite qui ne l'est pas. En langue naturelle, l'implicite est inférable lorsqu'il est de nature linguistique, c'est-à-dire qu'il est « calculable » par le seul fait du sens : « De la phrase *Je suis allé consulter un rhumatologue*, on infère analytiquement (par le sens) que *Je suis allé consulter un médecin*. » (Martin 1996, 131)

Appliquant cette notion à un discours lexicographique tel que le FEW, nous dirons que l'implicite est inférable lorsqu'il est « calculable » par le seul fait de la grammaire et du lexique propres au dictionnaire.

2.6.3.1 Implicite inférable par la grammaire du FEW

On trouve, dans le FEW, des informations implicites qui sont inférables à partir de la *grammaire fewienne*. Par grammaire fewienne, nous entendons les règles qui régissent la syntaxe du discours fewien et, notamment, la hiérarchisation et la structuration des éléments qui le composent. L'implicite inférable par la grammaire demande au lecteur de connaître (1) la position où il se trouve au sein des niveaux structurels supérieurs (micro-, macro- ou superstructurels) ou (2) le contexte proche dans lequel se situe l'implicite. Dans le premier cas, le raisonnement est le suivant : « dans la super(macro/micro)structure du FEW, je suis à tel endroit ; j'en déduis donc que ... ». Par exemple, « dans la macrostructure, je suis dans la section des étymons d'origine latine, préromane ou grecque ; j'en déduis donc que la langue implicite de l'étymon est le latin » ou « dans la microstructure, je suis dans la section II ; j'en déduis donc que les lexèmes sont probablement des emprunts, inférence que je vérifie par la lecture du commentaire ». Dans le second cas, le raisonnement consiste à prendre en compte ce qui a été formulé précédemment. Par exemple, « dans le contexte qui précède, la dernière occurrence d'une catégorie grammaticale est *f.* ; j'en déduis que le lexème est féminin » ou « dans le contexte qui précède l'endroit où devrait se trouver une glose d'étymon, se trouve l'étymon *auf* (FEW 1, 174a) : j'en déduis que la glose implicite de l'étymon est *auf* également ».

L'information implicite inférable par la grammaire se trouve donc résolue dans le FEW. Elle est donnée par la *position* de l'élément ou de l'ensemble des éléments, soit dans un niveau structurel supérieur, soit par rapport à des éléments de même niveau structurel exprimés précédemment dans le discours. Concrètement,

- la position microstructurelle (insertion au sein d'un paragraphe ou ensemble de paragraphes marqués alphanumériquement ou caractérisés par un titre explicite) permet de connaître les particularités morphologiques, phonétiques, sémantiques, de transmission ou autres qui s'appliquent aux lexèmes présents ;
- la position macrostructurelle permet de résoudre la langue à laquelle appartient un étymon-vedette et l'identité du signataire d'un article ;
- la position superstructurelle indique l'origine connue ou inconnue d'un lexème ;
- la position dans la séquence des unités permet de résoudre les molécules ellipsées d'une unité minimale de traitement ;
- la position de la glose d'étymon juste après son étymon permet d'inférer la glose implicite d'un étymon à partir de ce dernier.

2.6.3.2 Implicite inférable à partir du lexique du FEW

On trouve également dans le FEW des informations implicites qui sont *inférables à partir du lexique fewien*. Par lexique fewien, nous entendons les termes techniques utilisés par la langue fewienne et définis dans le *Beiheft*, qui donne à la fois leur nature, leur signification et des informations complémentaires les concernant. Par exemple, *Gdf* est un sigle bibliographique qui signifie « Godefroy, F., *Dictionnaire de l'ancienne langue française* ; 10 vol. ; Paris 1880–1902 » ; *abearn.* est une étiquette géolinguistique qui signifie « ancien béarnais » et qui indique que le lexème provient soit du *Dictionnaire*

béarnais ancien et moderne de Lespy et Raymond (Montpellier 1887), soit du *Livre des syndics des Etats de Béarn* (Paris – Auch 1900).

La référence bibliographique d'un lexème est inférable de cette façon, par la liste des sources associées dans le *Beiheft* à son étiquette géolinguistique. De même, la molécule de la datation est inférable par la date donnée dans le *Beiheft* pour les sources citées dans la molécule de la référence bibliographique ou pour les sources correspondant à l'étiquette géolinguistique. Le détail des abréviations géolinguistiques et bibliographiques du FEW appartient également à cette catégorie d'implicite, puisqu'il est résolvable aussi par le recours au *Beiheft*.

Cet implicite inférable à partir du lexique fewien permet aux lecteurs avertis qui connaissent ce lexique de ne pas passer par le *Beiheft* : leurs connaissances lexicales de la langue fewienne leur permettent de résoudre directement l'implicite et de raccourcir ainsi leur itinéraire de lecture.

Pour un lecteur qui ne connaît pas le lexique fewien, l'implicite est inférable de cette manière uniquement lorsqu'il se trouve effectivement résolu dans le *Beiheft*. Deux problèmes peuvent survenir, qui empêchent l'inférence. Le premier est la non exhaustivité et l'incomplétude des données du *Beiheft*. Sa troisième édition (Complément 2010) pallie ce problème par une mise à jour des données (termes techniques et informations associées). Le second problème est l'existence de variantes pour un même terme technique, variantes qui ne sont pas relevées dans le *Beiheft* puisque les termes techniques du discours fewien sont censés être normalisés. De telles incohérences sont explicables dans le FEW : elles résultent de la masse de données traitées, de la durée de parution et de la multiplicité des rédacteurs. Dans la plupart des cas, un terme de forme semblable répertorié dans le *Beiheft* permet au lecteur d'établir une hypothèse de « lemmatisation » de la variante qui résolve l'implicite.

2.6.3.3 Implicite inférable à partir de la grammaire et du lexique fewiens

Certains cas ambigus nécessitent pour être résolus le recours à la fois à la grammaire et au lexique fewiens. Par exemple, « Chambon » est un terme technique qui, dans le *Beiheft*, est défini à deux reprises, soit comme étiquette géolinguistique signifiant « Le Chambon-le-Château (Lozère, Mende, Grandrieu) », soit comme sigle bibliographique signifiant « Le commerce de l'Amérique par Marseille [...], Avignon 1764 [Barb] ». Le choix entre ces deux significations dépend du contexte et de la position où se trouve le terme dans la syntaxe fewienne (en tête d'unité ou en fin d'unité). Par ailleurs, si le terme ne se trouve pas au sein d'une unité minimale de traitement, mais à la fin du commentaire de l'article, il ne s'agit d'aucune de ces deux solutions, mais du nom du rédacteur et ancien directeur du FEW Jean-Pierre Chambon. La résolution de l'implicite demande ici au lecteur de connaître à la fois le lexique du FEW et les règles syntaxiques qui régissent son discours.

2.6.3.4 Implicite non inférable à partir du FEW

L'implicite en langue naturelle n'est pas inférable (mais malgré tout décodable) lorsqu'il n'est pas de nature linguistique, mais de nature pragmatique, c'est-à-dire qu'il est indissociable de connaissances d'univers :

Ablt. — Nfr. *pochée* f. „contenu d'une louche" (provinz., Trév 1743–1771), neuch. Vaud id., Blon. *potšá*, aost. *potsá*, Thônes *pochá*, Vaux *pǫϑá*, Lant. *pœtšá̊*, Cr. *poϑyá*.. — Troyes *pochotte* „petite louche" Gr, Clairv. *poichotte*, Fim. *pœtšåt*, Fraize *pœtsăt* Horning 117, Bar. *pųtsăt*, Brotte *pwẹšǫt*, Châten. *poŭëtchate*, Abond. *pǫϑę́ta*, Vd'Ill. *potsĕta* „cuiller à écrémer" Luchsinger 37, Cr. *poϑę́ta* „petite cuiller à pot"; terrt. *potsetta* „chaudronnier". Nfr. *pochetée* „contenu d'une louche" (provinziell, Trév 1743–1771), argonn. *poch'teie*,

FIGURE 2.4 – Exemple de regroupements suffixaux non explicités (FEW 9, 176a, POPIA)

> L'implicite est pragmatique quand il est indissociable de connaissances d'univers. Une proposition peut être réinterprétée en situation : *Il a fait* (aux élections présidentielles) *un score honorable* peut signifier, selon le cas, qu'il a fait mieux que la précédente fois, ou bien qu'il a atteint la barre fatidique des 5%, ou encore qu'il est au second tour. De telles interprétations ne viennent pas d'un "calcul du sens" : elles sont aussi variables que les situations elles-mêmes et tributaires des connaissances du monde. On les appelle parfois des *sous-entendus*. (Martin 1996, 131)

Appliquées au discours lexicographique du FEW, les connaissances d'univers sont les connaissances phonétiques, morphologiques, philologiques etc. qui sont mobilisées pour lire et décoder le discours.

Dans l'extrait de la figure 2.4, les suffixes à la base des regroupements de lexèmes ne sont pas explicités, ni par un marqueur, ni par le commentaire en fin d'article. Seul un lecteur muni de connaissances phonétiques sera capable de les décoder. De même, une date n'est pas toujours associée à une source dans le *Beiheft* (par exemple, *Coucy* est explicité « Chastellain de Coucy » sans aucune autre information dans la première édition du *Beiheft*) : dans ce cas, seules les connaissances philologiques du lecteur peuvent l'aider à attribuer une datation à la source et, ce faisant, au lexème.

2.6.3.5 Divergences selon les utilisateurs

Cette analyse en termes d'« inférabilité » de l'implicite fewien permet de comprendre les difficultés inhérentes à chaque type de lecteur : le lecteur qui ne connaît pas la grammaire et le lexique fewien, mais est compétent en linguistique française et dialectale ne rencontrera pas les mêmes problèmes de décodage que le lecteur familier du FEW mais privé des connaissances extrafewiennes nécessaires à la compréhension de certaines informations. On comprend aussi pourquoi les réponses reçues au questionnaire en ce qui concerne l'utilité d'expliciter les abréviations n'étaient pas homogènes (→ 2.4.2) : elles dépendaient du degré de maîtrise qu'avait chaque personne interrogée du lexique fewien. Le besoin d'explicitation pour la catégorie des étudiants était, par exemple, clairement affirmé – et tout à fait justifié.

2.7 Conclusion

En tentant, dans ce chapitre, d'appréhender l'objet FEW du point de vue de l'utilisateur, nous avons perçu le pourquoi du souhait d'informatisation : les besoins exprimés par les utilisateurs, justifiés par le contenu du FEW, ne sont pas résolus par les itinéraires de consultation et de lecture permis par le discours lexicographique sous sa forme actuelle. Les difficultés d'accès aux données du FEW ont toujours été attribuées à la présentation condensée et hautement structurée du discours fewien ; notre analyse révèle que les difficultés proviennent également du fait que les utilisateurs veulent consulter le FEW dans une optique plus large que celle pour laquelle il a été conçu, en privilégiant une dimension thesaurus qui, pour Wartburg, était avant tout un moyen de construire et d'étayer ses étymologies en se fondant sur une base documentaire solide. Cette vision du FEW comme un thesaurus est contrebalancée par la conscience très nette, chez les utilisateurs avertis, que le résultat produit est essentiellement un recueil de monographies et que c'est dans la structure du discours monographique que se trouve l'essentiel de l'analyse fournie par le FEW. Une exploitation du FEW dans une dimension thesaurus présente dès lors le risque de dénaturer la cohérence du discours monographique. La communauté scientifique se trouve prise entre deux extrêmes, certains utilisateurs encourageant une informatisation du FEW qui permettrait une exploitation tous azimuts et d'autres rejetant toute informatisation par crainte des conséquences néfastes qu'elle pourrait induire sur le FEW et sur son utilisation.

La solution consiste selon nous à modéliser le FEW dans ses deux dimensions, de manière à répondre aux besoins des utilisateurs sans dénaturer la cohérence du discours fewien. Il s'agit de rendre plus aisés les itinéraires de consultation et de lecture actuels tout en ouvrant l'accès aux nouveaux itinéraires de consultation attendus par la communauté et en optimisant les itinéraires de lecture : consultation transversale, lecture hypertextuelle au sein d'un article (lien entre documentation et commentaire via le marquage alphanumérique, lien entre notes et appels de note) et hors article (renvois internes du FEW, liens avec le *Beiheft*, liens avec des ressources externes), mise à jour du contenu. La modélisation doit également rétablir la dimension thesaurus en résolvant l'implicite chaque fois que c'est possible, c'est-à-dire chaque fois qu'il est inférable : l'accessibilité des données peut être optimisée dans les limites permises par les particularités du discours fewien. Le défi d'une modélisation intelligente du FEW consiste à permettre les utilisations en dimension T tout en conservant, et en respectant, la dimension M.

Chapitre 3

Modélisation du discours fewien

3.1 Introduction

Toute modélisation d'un objet donné, quelles que soient ses qualités et son niveau de perfection, est le résultat d'une interprétation parmi d'autres possibles de cet objet, qui dépend du point de vue adopté et des objectifs visés. Notre point de vue est celui de l'utilisateur du FEW et notre objectif principal la consultation et la lecture de l'ouvrage selon les itinéraires définis au chapitre précédent (→ 2.7). Il serait toutefois erroné d'en déduire que la modélisation doit correspondre à la seule vision qu'ont du FEW ses utilisateurs. En réalité, trois paramètres, qui sont en même temps des contraintes, influencent le modèle que nous proposons ici : les structures de l'ouvrage, les besoins des utilisateurs et les possibilités d'automatisation.

Les structures du FEW. Le modèle doit refléter les structures de l'ouvrage, c'est-à-dire représenter correctement les informations qui s'y trouvent et la structuration de ces éléments. Il est important de remarquer qu'à l'opposé d'un ouvrage lexicographique rédigé à partir d'une structure préalablement fixée et d'un guide à suivre, le FEW pose le problème de l'interprétation après coup d'un texte dont on cherche des règles de rédaction et de lecture. La description des structures du FEW établie par Éva Buchi (1996) nous sert de point de départ et de référence constante. Cette analyse est complétée par celle que font de leur travail les rédacteurs du FEW, anciens et actuels, interrogés dans le cadre de cette thèse[1]. Ces analyses structurelles, parfois divergentes[2], présentent un point de vue interne au FEW et se placent du côté de sa production. Dans cette optique, une formalisation a été proposée par Nicolas Mazziotta, qui cherche à rendre compte du raisonnement et de la pensée sous-jacente au discours étymologique du FEW de façon parallèle à celui de l'ALW (cf. Boutier 2008 ; Mazziotta 2011 ; Mazziotta et Renders 2010). L'ALW et le FEW présentent en effet, du point de vue de leur

[1] Nous remercions tout particulièrement Marie-Guy Boutier, Éva Buchi, Jean-Paul Chauveau, Yan Greub, France Lagueunière, Anne-Christelle Matthey et Christel Nissille pour leurs éclaircissements précieux sur de nombreux points de détail.

[2] Notamment à propos des règles qui conditionnent l'apparition d'une référence bibliographique (→ 3.6.4).

structure et de la façon dont s'élabore un article, un certain nombre de points communs.

Les besoins des utilisateurs. La modélisation du FEW doit, dans ses applications, permettre la consultation et la lecture de l'ouvrage selon les itinéraires définis au chapitre précédent. Cette contrainte conduit, par rapport aux analyses structurelles internes au FEW, à quelques libertés et prises de position qui rendent le modèle plus proche de la vision qu'ont du FEW ses utilisateurs. Par ailleurs, cette contrainte présente une priorité plus haute que la précédente : les éléments du FEW, pour autant qu'ils existent, dont l'identification n'est pas nécessaire pour répondre aux besoins d'utilisation ne sont pas obligatoirement à intégrer dans le modèle (→ 3.3.3).

Les possibilités d'automatisation. Les types d'information du FEW qui auront été définis dans le modèle devront être identifiés dans l'ensemble des articles du FEW. Cette identification dans le texte électronique se fera, concrètement, sous forme de balisage XML. Le FEW étant une œuvre monumentale qui peut difficilement être traitée manuellement, le défi de cette thèse consiste dès lors à permettre l'automatisation de ce balisage[3]. L'implicite fewien (→ 2.6) pose dans ce cadre un problème majeur.

Ces trois contraintes ne sont pas toujours compatibles et obligent à effectuer des choix. Notre objectif principal est de répondre aux besoins des utilisateurs : cet objectif justifie à lui seul l'informatisation du FEW. Toutefois, il ne peut se faire aux dépens des structures inhérentes de l'ouvrage. Une solution consiste à appréhender les structures du point de vue de l'utilisateur : les deux dimensions, thesaurus et monographique, du FEW sont, dans cette optique, des guides pour la modélisation. Enfin, la nécessité d'automatiser l'application du modèle restreint le choix des possibles et empêche de concevoir un modèle qui serait parfait en théorie, mais non applicable en pratique.

Lorsqu'un modèle théorique est appliqué sur un objet linguistique, et plus particulièrement lorsque cette application est informatique, il est nécessaire de représenter ce modèle dans un langage formel. Notre modèle étant destiné à être appliqué dans le texte électronique du FEW sous la forme de balises XML, nous le formaliserons directement dans ce langage. Les raisons et les principes qui guident ce choix sont expliqués plus loin (→ 3.4.1). Nous supposons que le lecteur de cette thèse est familiarisé avec le langage XML[4].

Ce chapitre se divise en sept parties. Nous commençons par définir les unités de base du modèle (→ 3.2). Nous expliquons ensuite les principes que nous suivons dans sa construction (→ 3.3) et nous justifions le choix du formalisme XML (→ 3.4). Nous construisons ensuite le modèle sur trois niveaux (→ 3.5–3.7). Enfin, nous synthétisons le modèle construit et nous donnons un exemple d'application sur un article du FEW (→ 3.8), avant de conclure (→ 3.9).

[3]Cette contrainte ne frappe pas le formalisme décrit dans Mazziotta 2011, qui a été proposé dans le cadre de la rédaction informatisée des nouveaux articles de la refonte et qui est destiné à être appliqué à ces articles de façon manuelle.

[4]Pour une introduction rapide à XML, voir `http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SG.html`.

3.2 Unités de traitement

La première question que nous nous posons, avant de construire un modèle, est de déterminer quel en sera l'élément central. Dans le cas qui nous occupe, à savoir la modélisation d'un discours lexicographique, cette question revient à déterminer quelle information constitue le centre d'intérêt du discours.

Dans le FEW, ce centre est double. Il dépend du point de vue qu'on porte sur l'ouvrage : le considère-t-on avant tout comme un recueil de monographies ou comme un thesaurus (→ 2.5) ? Dans une modélisation du FEW en dimension monographique, l'élément central est l'article, à considérer d'une part comme une unité autonome par rapport aux éléments structurels supérieurs (sections linguistiques ou ensembles conceptuels) et d'autre part comme une unité complexe non réductible à la simple énumération des entités qui le composent. L'article constitue l'*unité maximale de traitement* (Büchi et Chambon 1995, 948). Dans la vision du FEW comme thesaurus, en revanche, l'élément central est le lexème accompagné des informations moléculaires qui s'y rapportent directement, le tout formant ce que Büchi 1996 nomme l'*unité minimale de traitement*. Cette unité est également autonome et non réductible à la simple succession de ses composants. Elle correspond, en production, à la *fiche FEW*[5] (fiche élaborée par les rédacteurs lors du dépouillement des matériaux et utilisée ensuite comme base de travail lors de la rédaction de l'article) et, en réception, à l'unité informationnelle qui forme la cible des recherches des utilisateurs en dimension T. Afin de ne pas la confondre avec le lexème vu comme signe linguistique, nous l'appellerons dorénavant la *cellule lexicale*.

Les deux unités de traitement que sont l'article et la cellule lexicale correspondent aux deux unités de consultation traditionnelles du FEW que sont l'étymon d'une part, le lexème d'autre part[6]. Etymon et lexème ne sont pas indépendants, mais liés par une relation verticale que nous avons déjà schématisée (→ 2.3.1) et qui signifie « provient de » ou « est à l'origine de » selon le sens descendant ou ascendant du parcours.

Dans les faits, ce schéma est cependant trop simpliste, car le lexème peut se rattacher à l'étymon via des étymons intermédiaires (sous-lemmes d'un point de vue lexicographique) ou des lexèmes intermédiaires, suivant une histoire parfois très complexe. En pratique, ces relations complexes sont rendues dans le FEW, d'abord par la structuration de l'article en plusieurs groupements hiérarchisés de lexèmes, ensuite par la simple succession des cellules lexicales, qui montre par exemple qu'un lexème dérive de celui qui précède (cf. Mazziotta et Renders 2010).

Une modélisation du FEW selon la dimension M peut donc s'élaborer en plusieurs étages. L'étage supérieur est représenté par l'article (unité maximale de traitement) et l'étage inférieur par la cellule lexicale (unité minimale de traitement). Ces deux étages sont reliés par un ensemble d'étages intermédiaires qui opèrent des regroupements de plus en plus fins entre les cellules lexicales. Par exemple, la structure de l'article ĬNCHOARE (FEW 4, 622b-623a), comprenant 16 lexèmes (L), contient trois étages intermédiaires, si l'on prend en compte uniquement les groupements de lexèmes explicités par le marquage alphanumérique (cf. figure 3.1).

[5]Cf. Chambon 1989a, 220.

[6]L'étymon est bien entendu un lexème lui aussi, mais nous réservons le terme de lexème aux unités lexicales galloromanes dont le FEW explique l'étymologie : les *explicanda*. L'étymon est considéré dans ces pages comme le représentant de la famille lexicale expliquée dans un article : l'*explicans* (→ 2.3.1).

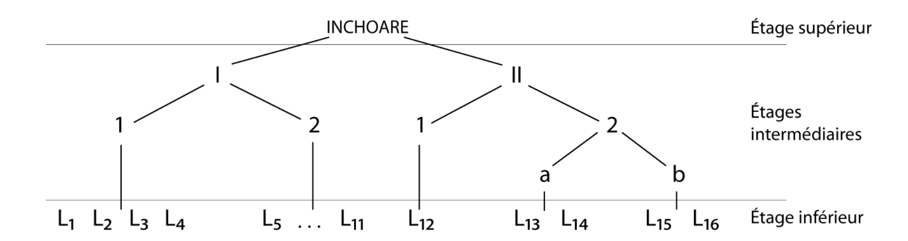


FIGURE 3.1 – Structure de l'article ĬNCHOARE (FEW 4, 622b-623a)

Une modélisation du FEW selon la dimension T, en revanche, a pour étage supérieur la cellule lexicale (qui devient alors l'unité maximale de traitement) et pour étage directement inférieur les molécules qui la composent. Ces molécules sont liées à la cellule lexicale par des relations de détermination, tout à fait semblables à celles que Marie-Guy Boutier a définies dans l'ALW : détermination de catégorisation, de signification, de localisation, d'origine et déterminations accessoires (cf. Boutier 2008), auxquelles il faut ajouter pour le FEW une détermination de datation, ainsi qu'une détermination de forme (le signifiant étant une molécule comme les autres).

L'article et la cellule lexicale constituent chacun l'unité supérieure respectivement dans la dimension M et dans la dimension T du FEW. On pourrait parler en dimension M de nomenclature étymologique (l'article étant symbolisé par un étymon) et, en dimension T, de nomenclature lexicale.

3.3 Principes de base

La construction du modèle à partir des deux unités de base que sont l'article et la cellule lexicale suit quelques principes que nous exposons ici. Il s'agit des principes de redescente des informations, de limitation dans la résolution de l'implicite et de limitation dans l'identification des informations.

3.3.1 Redescente des informations

Le fait que l'article (dans la dimension M) et la cellule lexicale (dans la dimension T) constituent des unités supérieures implique qu'elles doivent être autonomes, c'est-à-dire que même sorties de leur contexte, elles doivent contenir toutes les informations qui les concernent directement. Prenons pour exemple l'article ELF (FEW 18, 59a), où la langue à laquelle appartient l'étymon est absente en structure de surface. En structure profonde, elle est déductible de la position de l'article dans le volume 18, qui contient les matériaux d'origine anglaise. Deux solutions s'offrent à nous pour prendre en compte cette information macrostructurelle dans la modélisation du FEW. La première consiste à inclure les articles dans un élément supérieur qui serait la section linguistique de l'étymon. L'article perd alors son autonomie, car cette information qui le concerne directement n'est plus accessible une fois que l'article est retiré de son contexte. La seconde solution consiste à redescendre l'information au niveau de l'article, en rétablissant la langue de l'étymon dans le champ de l'entrée. Cette solution rétablit l'autonomie de l'article : l'information macrostructurelle en fait dès lors partie intégrante.

parfaire „opérer, obtenir, créer“ Destrees, „terminer complètement“ (Sens 1555, CoutGén 3,516). — Redensarten Nfr. *parfaire le juste prix* „réparer la lésion qu'a éprouvée le vendeur d'un immeuble“ (seit Ac 1835); mfr. nfr. *faire et parfaire* „achever complètement“ (Est 1538–Stœr 1625), *faire et parfaire un procès* „conduire un procès criminel jusqu'au jugement dernier“ (1522–Land 1851, CoutGen 3, 1280; D'Aubigné), *parfaire un procès* (Pom 1671–Trév 1771); *fait et parfait* „entièrement terminé (d'un ouvrage)“ (seit 1443).

FIGURE 3.2 – Locutions (FEW 8, 238b, PĔRFĬCERE)

Ce principe de redescente de l'information s'applique de la même façon à la cellule lexicale qui forme l'unité supérieure en dimension T. Prenons comme exemple les cellules lexicales regroupées entre la troisième et la dernière ligne de l'extrait de la figure 3.2, qui présentent un point commun : ce sont des locutions.

La mention *Redensarten* qui l'indique se situe en dehors de la cellule lexicale, dans un niveau structurel supérieur. Dans une utilisation du FEW en dimension M, cette information est accessible, car donnée par les étages intermédiaires qui expliquent les relations entre l'étymon et le lexème (→ 3.2). Dans une utilisation du FEW en dimension T, où les cellules lexicales sont appréhendées de façon autonome, cette information morphologique devient peu accessible : une consultation transversale du FEW selon ce critère, ou une mise à jour de l'ouvrage nécessitant un déplacement d'un des lexèmes, requièrent que cette information soit étroitement liée aux lexèmes concernés. Redescendre les critères de regroupement microstructurels au niveau de la cellule lexicale permet d'assurer son autonomie[7].

Exprimé de façon plus formelle, le principe de redescente des informations consiste à attribuer à des *objets* (ou individus) une propriété qui dans le FEW est située au niveau d'une *classe* d'objets (ou groupement d'individus). Cette redescente d'une information de la classe à l'objet est à la fois un déplacement de l'information, permettant de réduire la distance entre celle-ci et les éléments qu'elle explique, et une reproduction de l'information, effectuée autant de fois qu'il existe d'éléments concernés. Le résultat conduit à rendre l'information plus accessible et à résoudre l'implicite présent dans le discours fewien.

3.3.2 Limitations à la résolution de l'implicite

Nous avons tenté une analyse de l'implicite fewien du point de vue du lecteur humain (→ 2.6). Or, une des contraintes qui s'exerce sur notre modélisation est de la rendre automatisable. Il est dès lors nécessaire de s'interroger sur les possibilités de résolution de cet implicite par une machine. Nous avons défini un algorithme comme une succession de règles qui, appliquées de façon aveugle, mènent au résultat recherché. Tout implicite est donc résolvable algorithmiquement si l'on peut définir, pour sa résolution, un ensemble de règles strictes à suivre méthodiquement. Reprenons la typologie établie précédemment (→ 2.6.3) concernant le décodage de l'implicite fewien par un humain.

[7] La rédaction modulaire des articles de la refonte suit le même principe, puisque les critères qui serviront à regrouper, classer et hiérarchiser les lexèmes dans l'article sont d'abord associés à chaque lexème de façon individuelle (cf. Matthey et Nissille 2010).

3.3.2.1 Implicite de nature grammaticale

Les informations implicites inférables par la grammaire fewienne sont décodables par un algorithme (1) si cette grammaire a préalablement été donnée à l'algorithme et (2) si la position de l'information dans le FEW est connue de l'algorithme. Par exemple, la langue à laquelle appartient un étymon-vedette est décodable si l'algorithme dispose des langues implicites correspondant à chaque section linguistique et qu'il sait dans laquelle de ces sections se trouve l'étymon. En pratique, il est généralement possible de définir des règles "grammaticales" utilisables par un algorithme à partir du moment où ces règles existent pour le lecteur humain.

3.3.2.2 Implicite de nature lexicale

Les informations dont l'accessibilité dépend de connaissances lexicales sont décodables par un algorithme à condition qu'il dispose du lexique requis. Dans le cas du FEW, la résolution de ce que nous nommons l'implicite lexical (→ 2.6.3.2) dépend donc des modalités d'informatisation du *Beiheft*, ainsi que de sa mise à jour.

3.3.2.3 Implicite de nature pragmatique

L'implicite de nature pragmatique requiert la mobilisation de connaissances encyclopédiques, c'est-à-dire de connaissances linguistiques, philologiques, phonétiques etc. qui ne sont pas exposées dans le FEW. A priori, cet implicite ne peut être résolu algorithmiquement, puisqu'il demande réflexion de la part d'un lecteur humain.

Dans le cas du FEW, certaines de ces connaissances (surtout philologiques) sont néanmoins accessibles via le *Beiheft* et se rapprochent de l'implicite lexical (→ 2.6.3.4) puisqu'elles deviennent alors inférables. Cet implicite est dès lors résolvable par un algorithme à condition qu'il dispose, sous une forme utilisable, du réservoir de connaissances encyclopédiques contenu dans le *Beiheft* : comme pour l'implicite lexical, les modalités d'informatisation et la mise à jour de ce complément au FEW jouent ici un rôle déterminant.

3.3.2.4 Traitement automatique de l'implicite

De ce qui précède, il s'ensuit que, pour qu'un algorithme puisse rétablir une information implicite en structure de surface, trois conditions sont nécessaires. Il faut tout d'abord (1) que l'information soit inférable, c'est-à-dire qu'elle apparaisse explicitement quelque part dans le dictionnaire (FEW ou *Beiheft*), quelle que soit la distance entre cet endroit et celui où elle est implicite. En d'autres termes, il faut qu'elle appartienne à la structure profonde du dictionnaire. Il est ensuite nécessaire (2) que cet endroit distant où se trouve l'information soit connu de l'algorithme et (3) que l'information puisse être extraite de cet endroit.

Parmi les trois catégories d'implicite mentionnées ci-dessus, les deux premières posent peu de problèmes. L'implicite de nature grammaticale est résolu au sein du FEW et est accessible : il répond généralement à des règles strictes qui peuvent être exprimées de façon algorithmique. L'implicite de nature lexicale dépend quant à lui d'une correspondance terme à terme qui peut facilement être extraite du *Beiheft*. En

revanche, la troisième catégorie pose problème. L'extraction de connaissances encyclopédiques à partir du *Beiheft* est possible uniquement si ce dernier est informatisé de façon à expliciter clairement les informations implicites (notamment de datation et de localisation) cachées derrière chaque sigle. L'informatisation du *Beiheft* nécessite dès lors une étude philologique approfondie de chacune des références bibliographiques qui y sont consignées, étude qui sort du cadre de cette thèse. Il n'est en outre pas du tout certain que chaque sigle puisse être associé à une datation et à une localisation : ces données sont, dans certains cas, soit sujettes à révision en fonction de l'avancée des études philologiques, soit impossibles à définir (par exemple pour des textes médiévaux). L'accessibilité des informations gérées par cet implicite pragmatique est donc trop incertaine, au moment où nous écrivons ces lignes, pour que celui-ci puisse être pris en compte dans notre modèle, malgré son intérêt évident. De même, la modélisation proposée ici ne pourra intégrer des éléments dont l'explicitation mobilise des connaissances encyclopédiques non accessibles via le *Beiheft* et qui nécessitent dès lors une réflexion et une décision de la part d'un lecteur humain.

En résumé, nous considérons que les algorithmes disposent d'une connaissance grammaticale et lexicale du FEW. Tout implicite n'entrant pas dans une de ces deux catégories sera considéré, dans le cadre de la modélisation proposée ici, comme non résolvable algorithmiquement[8]. La redescente des informations implicites est dès lors possible uniquement si l'information est inférable grammaticalement ou lexicalement, ou, en d'autres termes, si l'information se trouve effectivement quelque part dans le FEW ou dans le *Beiheft* et qu'elle peut être retrouvée en suivant un ensemble de règles.

3.3.3 Limitation à l'identification des informations

Le fait que l'implicite de nature pragmatique ne soit pas automatisable et, donc, ne puisse être intégré au modèle ne pose pas de problème majeur pour la construction de ce dernier. Cet état de fait est possible parce que le modèle ne cherche pas à être exhaustif. Ce choix pourrait être considéré comme un défaut, notamment en comparaison avec les modèles qui dirigent la rédaction des dictionnaires actuels. Dans un dictionnaire élaboré directement sur ordinateur, tous les types d'information sont, en effet, définis et prévus dans le modèle qui guide la rédaction (ou « grammaire du dictionnaire », cf. Kilgariff 2005, 98). Cela signifie que chaque élément du dictionnaire trouve sa place dans un champ informationnel, ou, en d'autres termes, que chaque élément est identifié. Dans le cas d'une rétroconversion, il serait tentant de chercher également à identifier la totalité des types d'information contenus dans le dictionnaire imprimé, de façon à réécrire en quelque sorte le modèle tel qu'il aurait pu exister comme guide lors de la rédaction. Malheureusement, la rédaction du FEW s'est faite sur une longue période, par de multiples rédacteurs, et avec des changements de méthode importants qui ont laissé des traces dans la structure du dictionnaire. Malgré l'existence d'un squelette structurel commun à tous les articles (mis au jour par Büchi 1996) et de règles de rédaction rigoureuses, les dimensions de l'ouvrage ne permettent pas de maîtriser la totalité des incohérences et des erreurs qui s'y cachent. Il faut se résoudre à cette constatation : un dictionnaire tel que le FEW ne permet pas de construire un modèle

[8]Rappelons que cette limitation au traitement de l'implicite, due à la contrainte d'automatisation, n'est pas définitive, puisque le modèle sous-jacent à la rétroconversion du FEW pourra être enrichi par la suite de façon à intégrer manuellement davantage d'information, dont l'implicite pragmatique (cf. Mazziotta et Renders 2010).

exhaustif, qui couvrirait l'ensemble des éléments qu'il contient en structure de surface et qui prendrait en compte toutes les structurations possibles de ces éléments.

Heureusement, tel n'est pas notre but. Nous tentons d'identifier uniquement les types d'information qui répondent aux trois contraintes définies ci-dessus (→ 3.1) et, surtout, qui permettent les itinéraires de consultation et de lecture souhaités par les utilisateurs. De même que l'identification de l'implicite était limitée par les possibilités d'automatisation, nous nous permettons de limiter l'identification des types d'information du FEW à ceux qui ont été définis par Büchi 1996 et qui présentent une utilité pour les utilisateurs. Il n'est donc pas question de créer un modèle exhaustif : au contraire, il est permis de laisser certaines portions de texte non identifiées. Ce faisant, nous suivons un principe consistant à ne pas identifier de façon explicite ce qui n'a pas besoin d'être identifié. Nous évitons la tentation de modéliser un maximum d'informations, pour nous contenter de ce qui est pertinent.

3.4 Formalisation en XML

3.4.1 Choix du formalisme XML

Il est important de distinguer modélisation et formalisation. Un modèle est une construction/description d'un objet, qui peut se concevoir de façon théorique, indépendamment de tout formalisme. En pratique, il est toutefois nécessaire, à un moment donné, de le représenter sous une forme concrète. La formalisation choisie doit être adaptée aux caractéristiques du modèle et doit permettre de représenter toutes les propriétés définies dans le modèle.

Büchi 1996, qui décrit les structures du FEW, construit un modèle théorique qui n'est associé à aucun formalisme. Cela ne pose aucun problème, car cet ouvrage reste dans le domaine théorique. La modélisation que nous proposons ici, en revanche, est destinée à être appliquée dans le cadre d'une informatisation de l'ouvrage. Cet objectif concret oblige à formaliser le modèle dans un langage adapté à cet objectif.

À vrai dire, le choix du formalisme XML s'est imposé à nous dès le début du projet, pour des raisons au départ externes. Le langage XML est devenu un standard utilisé dans la plupart des projets linguistiques actuels touchant à l'informatisation de documents textuels. XML apparaît comme « la syntaxe de document la plus robuste, la plus fiable et la plus flexible jamais inventée »[9]. L'analyse des propriétés de ce langage nous a convaincue que, malgré des limitations qu'il ne faut pas oublier (dues notamment à la structure en arbre qu'elle impose), une formalisation en XML convenait à la modélisation du FEW que nous cherchions à construire. Nous sommes persuadée en effet que cette première étape de l'informatisation qu'est la rétroconversion ne nécessite pas un formalisme trop complexe, mais qu'il faut au contraire privilégier un formalisme qui soit le plus simple possible et qui permette ensuite amélioration et enrichissement sur une base solide. Or, un grand avantage de ce format universel de transmission des données textuelles est qu'il impose au document une structure, sans pour autant la prédéfinir ni en brider l'évolution : nous pouvons créer les éléments que nous voulons et les adapter à nos besoins. Dans le cas du FEW, ce confort est avant tout une nécessité. Ce standard a également l'avantage d'être indépendant de toute application informatique,

[9]Harold et Means 2000, xiii.

ce qui assure la conservation du document. Une fois le FEW balisé, nous pourrons décider de le traiter d'une façon ou d'une autre, y compris transformer le balisage pour qu'il réponde à de nouveaux objectifs.

3.4.2 Usage raisonné de la syntaxe XML

L'usage de la syntaxe XML nous permet d'insérer dans le texte du FEW toutes les informations que nous voulons sous la forme de balises et d'attributs. Ce formalisme est intéressant pour nous, car nous voulons conserver le reflet du FEW tel qu'il se présente dans son inscription traditionnelle et nous cherchons dès lors à modifier le moins possible la structure de surface du dictionnaire. Dans cette optique, nous définissons un principe de travail rigoureux et non négociable : toute analyse du FEW effectuée dans la modélisation, qu'il s'agisse d'identifier un type d'information ou de rétablir une information implicite, sera ajoutée sous forme de balise ou d'attribut uniquement. Le texte original du FEW sera quant à lui complètement disponible, de façon intégrale et sans modification aucune, en dehors des balises. Cela signifie que tout contenu textuel situé entre une balise XML ouvrante et une balise XML fermante représente un extrait de texte du FEW, tel qu'il se présente dans son inscription traditionnelle, et que les informations que nous portons sur ce contenu sont confinées à l'intérieur des balises mêmes. Ce principe est valable uniquement dans le cadre de la rétroconversion et ne s'applique pas à la phase d'exploitation des articles rétroconvertis (→ 4.1). En effet, l'exploitation des articles nécessitera très certainement des ajustements du formalisme XML afin d'optimiser les opérations effectuées par le moteur de recherche. Lorsque le formalisme défini est susceptible de modifications futures dans ce but, nous le signalerons en note.

3.4.3 Formalisation des principes de base

En pratique, tout type d'information défini dans le modèle sera identifié au moyen d'un élément XML. La question se pose de savoir comment utiliser adéquatement les balises et les attributs pour exprimer le modèle en suivant les principes définis plus haut (→ 3.3). Le principe de limitation à l'identification des informations ne pose aucun problème : un type d'information non retenu dans la modélisation ne sera simplement pas balisé. En ce qui concerne les principes de redescente des informations et de résolution de l'implicite, le langage XML apporte deux solutions, qui ne sont pas exclusives, et qui correspondent aux deux dimensions du FEW. Toutes deux utilisent les attributs associés aux éléments XML. Les attributs permettent d'expliciter ce que nous voulons sous la forme de paires nom-valeur : un attribut nommé volume, par exemple, pourrait recevoir la valeur « 1 », « 2 », « 3 » etc. selon le numéro de volume concerné. L'attribut se place à l'intérieur de la balise ouvrante de l'élément concerné, comme ceci : <few volume= « 1 »>.

Une première solution pour redescendre une information consiste à la placer comme attribut de l'élément où nous voulons qu'elle soit explicitée. Par exemple, nous pourrions associer à chaque cellule lexicale, identifiée par un élément XML (par exemple <cellule>), des attributs contenant toutes les informations utiles qui concernent cette cellule pour une utilisation du FEW en mode Thesaurus. Dans l'exemple cité plus haut de l'article PĔRFĬCERE, chacune des cellules rangées après le titre *Redensarten* pourrait par exemple recevoir un attribut indiquant qu'il s'agit d'une locution, comme ceci :

<cellule type = « locution »>. Cette solution de réexplicitation correspond exactement au principe de redescente des informations tel que nous l'avons défini. Elle est cependant, dans certains cas, très verbeuse et donc coûteuse. Elle ne doit donc être appliquée que là où elle est nécessaire, en fonction des besoins des utilisateurs.

Une autre solution consiste à profiter de l'imbrication d'éléments XML pour que tous les éléments X situés dans un même élément parent Y reçoivent la propriété donnée en attribut de cet élément Y. Nous pouvons par exemple décider que les cellules lexicales incluses dans un élément Y héritent des attributs de cet élément Y, sans qu'il soit nécessaire de créer un attribut dans chaque élément <cellule>. Cette solution reflète la dimension monographique du FEW : elle ne redescend pas les informations, du moins pas au niveau du formalisme. La redescente des informations est toutefois rendue possible lors de l'étape d'exploitation du document XML, par le biais de règles précises données au moteur de recherche qui traitera les demandes des utilisateurs. Par exemple, une consultation visant à relever toutes les locutions du FEW pourrait consister à chercher tous les éléments <cellule> qui seraient compris dans un élément supérieur Y contenant l'attribut type= « locutions ». Dans ce cas, il faut évidemment que l'information « locutions » ait été explicitée dans l'élément Y. Cette explicitation au niveau de l'élément Y peut se faire soit dans un attribut, soit dans le nom même de l'élément Y (par exemple <locutions>).

Nous possédons donc deux moyens de redescendre une information en utilisant le formalisme XML : (1) par la présence d'attributs explicitant l'information au niveau désiré (ce qui privilégie la dimension thesaurus) et (2) par l'imbrication des éléments XML (ce qui reflète la dimension monographique du FEW). Nous adoptons pour principe de ne pas multiplier les éléments XML et de privilégier la redescente des informations en tant qu'attributs au niveau des unités de base chaque fois que cela peut se faire sans alourdir le balisage de façon excessive. La seconde solution est privilégiée uniquement lorsque l'application de la première se révèle trop verbeuse.

3.5 Modélisation de l'étage supérieur : l'article

Nous examinerons ici ce qui concerne l'étage supérieur, représenté par l'article (unité maximale de traitement en dimension M). Seront traitées successivement ci-dessous la redescente des informations situées aux niveaux macrostructurel et superstructurel du FEW, la délimitation de l'article dans les matériaux d'origine inconnue, l'identification des relations entre articles (renvois internes) et la modélisation du champ de l'entrée.

3.5.1 Redescente des informations macro- et superstructurelles

La décision de considérer l'article comme l'unité maximale de traitement signifie pour la modélisation que le FEW est directement composé d'articles, sans éléments intermédiaires. Si nous formalisons en XML cette modélisation, que nous représentons l'ensemble des 25 volumes du FEW par l'élément <few> et un article par l'élément XML <art>, cette décision signifie que l'élément <few> contient uniquement des éléments <art>.

Nous exprimons cette propriété de la façon suivante[10] :

[10]Nous renvoyons à Van der Vlist 2002 pour une aide à la lecture des schémas XML utilisés ici.

```
<xs:element name="few" type="few" />
    <xs:complexType name="few">
    <xs:sequence minOccurs="1" maxOccurs="unbounded">
        <xs:element name="art" type="art" />
    </xs:sequence>
</xs:complexType>
```

Cet extrait de schéma XML peut être lu ainsi : il existe un élément nommé <few> (élément racine de notre modèle[11]), qui est composé d'une séquence d'éléments nommés <art>. Cette séquence comprend au minimum un élément <art> et au maximum une infinité d'éléments <art>.

Le modèle n'identifie donc dans le FEW aucun élément de niveau supérieur à l'article. Or, le regroupement des articles de la partie étymologisée en cinq grandes sections linguistiques (Büchi 1996, 24) est une information à conserver, notamment pour permettre des consultations tranversales sur le critère de la langue d'appartenance d'un étymon. La séparation des matériaux selon leur statut étymologique (matériaux étymologisés *vs* matériaux d'origine inconnue) est également importante, notamment pour la mise à jour de l'ouvrage. Chaque article du FEW étant considéré comme une unité de base autonome (et non comme un sous-élément d'une section linguistique ou conceptuelle), la redescente de ces informations macrostructurelles et superstructurelles au niveau de l'article s'impose.

Il suffit pour ce faire de *situer* l'article à un endroit précis du dictionnaire, ce qui revient à l'identifier au moyen de son *adresse FEW*. Le volume, la page et la colonne correspondant au début de l'article suffisent en effet, avec la mention de l'étymon-vedette, pour référencer un article de façon unique[12] et pour le situer à la fois dans la macrostructure et dans la superstructure du FEW, démarche qui est indispensable pour les inférences liées à la grammaire fewienne.

À chaque élément article (<art>) seront donc associés le volume, le tome éventuel, la page et la colonne où se situe son étymon-vedette. Suivant les principes adoptés pour la formalisation (→ 3.4.3), ces informations sont fournies en tant qu'attributs XML, idéalement contenus dans la balise ouvrante <art>[13], comme ceci :

<art book="1" volume="24" pg="34" s="a">

On pourrait penser que la mention des volume, tome, page et colonne où commence l'article, associée à la mention de l'étymon-vedette, renseigne également l'ordre d'apparition des articles dans le FEW. En réalité, l'ordre dans lequel sont classés les étymons à l'intérieur d'une section linguistique n'est pas toujours alphabétique (cf. Büchi

[11]L'élément racine, en terminologie XML, est l'élément duquel dépendent directement ou indirectement tous les autres éléments définis dans le document.

[12]La mention seule de l'étymon-vedette ne suffit pas car il existe des doublons (voir par exemple l'étymon CHOCOLATL, FEW 2, 642b et FEW 20, 63b).

[13]En pratique, lors de l'implémentation des algorithmes de rétroconversion, le souci de cohérence avec la balise <col/> indiquant les sauts de colonne et de page a conduit à préférer l'utilisation de la même balise <col/> pour signifier la page et la colonne où se situe le début de l'article. Dans le résultat de la rétroconversion, la formalisation de l'adresse FEW est donc finalement répartie sur deux balises qui se succèdent directement :

<art book="1" volume="24">
<col pg="34" s="a"/>

Cette formalisation est moins juste en termes de sémantisme des balises (la balise <col/> étant censée indiquer exclusivement un saut de colonne effectif dans le FEW), mais facilite le processus de rétroconversion et ne pose aucun problème pour l'exploitation du FEW rétroconverti.

1996, 23). Cela signifie que l'adresse FEW telle que définie ci-dessus ne suffit pas toujours pour connaître l'ordre exact dans lequel apparaissent deux articles débutant dans une même colonne. Or, cette information peut être intéressante à conserver, non pour l'analyse des lexèmes, mais dans une perspective de rendu exact à l'écran de la version originale du FEW. Il peut en outre arriver que des étymons homonymes se suivent dans la même page du FEW, ce qui rend l'adresse FEW non suffisante pour référencer un article de façon unique (cf. l'exemple des deux articles AHNE [FEW 15/1, 9ab] relevé par Büchi 1996, 85). Afin de conserver l'ordre séquentiel des articles, nous complétons l'adresse FEW en associant également à chaque article un numéro d'ordre dans la colonne, donné par l'attribut *ici* (*in-column index*).

Nous y ajoutons un attribut *id* qui donne à chaque article un numéro l'identifiant de façon unique. Cet identifiant est ajouté pour des raisons informatiques uniquement et ne présente aucune signification particulière par rapport au FEW.

La définition des attributs présents dans la balise ouvrante <art> s'effectue de la façon suivante :

```
<xs:complexType name="art">
    <xs:attribute name="id" type="xs:nonNegativeInteger" use="required"/>
    <xs:attribute name="ici" type="xs:positiveInteger" use="required"/>
    <xs:attribute name="volume" type="volume-id" use="required"/>
    <xs:attribute name="book" type="book-id" use="required"/>
    <xs:attribute name="pg" type="xs:positiveInteger" use="required "/>
    <xs:attribute name="s" use="required">
        <xs:simpleType>
            <xs:restriction base="xs:string">
                <xs:enumeration value="a"/>
                <xs:enumeration value="b"/>
            </xs:restriction>
        </xs:simpleType>
    </xs:attribute>
</xs:complexType>
```

Cet extrait de schéma XML dit qu'il existe un élément <art> qui comporte six attributs, nommés *id*, *ici*, *volume*, *book*, *pg* et *s*. Les valeurs possibles pour chacun de ces attributs sont définies dans le champ *type*. En particulier, l'attribut *s* est défini de façon à n'accepter que les valeurs « a » et « b » (correspondant à l'indication de la colonne dans laquelle se trouve le début de l'article).

3.5.2 Traitement des matériaux d'origine inconnue

Les matériaux d'origine inconnue (volumes 21 à 23) posent un problème de modélisation quant à la question de savoir où commence et où se termine un article. Büchi 1996 a montré qu'en structure profonde, les articles de cette partie superstructurelle ne se différencient pas des articles de la partie étymologisée, à condition de considérer le champ de l'entrée comme un module facultatif (cf. Büchi 1996, 76-78). Les concepts, malgré leur présentation typographique en grasses qui les fait ressembler à des lemmes, se situent à un niveau macrostructurel et représentent uniquement un moyen de classement des articles en l'absence d'un étymon. Sous chaque concept peuvent dès lors se trouver plusieurs articles proprement dits.

L'exemple cité par Büchi 1996 de « cambouis » (23, 75b-6a) montre que la reconnaissance d'un article suivant cette analyse est toutefois moins évidente que dans la partie étymologisée, car elle ne dépend pas de la typographie, mais est soumise à un examen attentif des types lexicaux. Cette propriété est problématique dans le cadre d'une modélisation qui doit pouvoir déboucher sur un traitement automatique. Or, la délimitation de l'élément article est essentielle, puisque ce dernier forme l'unité supérieure du modèle. Une solution provisoire consiste à considérer que dans la partie dite des Inconnus, un article comprend tous les matériaux regroupés sous un concept. Cette solution est en accord avec l'intuition du lecteur ; elle reflète la structure de surface du FEW et facilite l'automatisation du modèle. Un traitement ultérieur, manuel, pourra éventuellement être envisagé pour rétablir la structure profonde.

Afin d'identifier ces articles particuliers, nous ajoutons à l'élément XML <art> la possibilité de recevoir un attribut *type*, dont la valeur sera "concept" dans le cas des articles de la partie des Inconnus :

```
<xs:attribute name="type" type="article-type"/>
<xs:simpleType name="article-type">
    <xs:restriction base="xs:string">
        <xs:enumeration value="concept"/>
    </xs:restriction>
</xs:simpleType>
```

3.5.3 Le champ de l'entrée

3.5.3.1 L'entrée dans les matériaux étymologisés

Büchi 1996, 79 identifie dans le champ de l'entrée trois modules : l'étymon-vedette, la langue de l'étymon et la glose explicative de l'étymon, les deux derniers modules étant facultatifs.

L'étymon-vedette, représentant de l'article et unité de consultation, joue évidemment un rôle capital dans la modélisation du FEW. Son identification est essentielle dans les itinéraires de consultation traditionnels. Elle permet, dans un dictionnaire informatisé, de faire remonter à la surface les cas où l'étymon est le lemme de deux articles différents du dictionnaire (voir par exemple CHOCOLATL, FEW 2, 642b et FEW 20, 63b). Par ailleurs, le balisage des étymons-vedettes est nécessaire pour permettre les itinéraires hypertextuels entre articles via les « renvois FEW » (→ 3.5.4). Nous spécifions pour chaque étymon, formalisé par l'élément XML <etymon>, un *type* qui recevra la valeur « vedette » lorsque l'étymon est le lemme de l'article dans lequel il se trouve.

La langue de l'étymon constitue également un type d'information très attendu dans le cadre de consultations transversales : son balisage permet de rechercher des lexèmes selon la langue à laquelle se rattache leur étymon. Par ailleurs, en dimension M, son identification permet de préciser la section linguistique de l'étymon donnée par l'adresse FEW de l'article (→ 2.3.2.3). En effet, le numéro de volume contenu dans cette dernière rend accessible la section linguistique générale dans laquelle se trouve l'article, mais il n'indique pas la langue exacte à laquelle se rattache l'étymon au sein de cette section. L'identification de la langue de l'étymon, que cette dernière soit explicite ou implicite dans l'inscription traditionnelle du FEW, est dès lors nécessaire autant

en dimension T qu'en dimension M.

La glose de l'étymon ne nécessite pas obligatoirement une identification, pour deux raisons au moins. La première est qu'elle ne répond à aucun besoin des utilisateurs. En dimension T, la glose est une indication qui n'est pas assez structurée pour permettre d'effectuer des recherches sémantiques dans le FEW. Des consultations transversales du FEW selon des critères sémantiques seront permises de façon plus efficace après la mise à disposition d'un index onomasiologique (projet en cours d'élaboration au Centre du FEW, → 1.5.1). Les seules gloses intéressantes pour les consultations transversales seraient les gloses du type *personenname*, *ON*, *NP*, ou encore *PN* (cf. Büchi 1996, 288), qui permettent d'identifier certains noms propres. Malheureusement, ces informations sont loin d'être systématiques. La liste des éponymes constituée par Büchi 1996 (564-589) permet plus efficacement l'identification de cette catégorie particulière d'étymons. La même remarque vaut pour l'identification des étymons onomatopéiques relevés par Büchi 1996 (392-393). La deuxième raison est que l'identification du champ de l'entrée, de l'étymon et de la langue de l'étymon suffisent pour rendre la glose accessible à l'utilisateur d'une version informatisée du FEW, puisqu'il lui suffit de considérer le texte contenu dans l'entrée qui n'est ni un étymon, ni une langue d'étymon. En n'incluant pas la glose de l'étymon dans le modèle, nous suivons le principe consistant à ne pas expliciter ce qui n'a pas besoin de l'être (→ 3.3.3).

Un quatrième élément peut apparaître, mais très rarement, dans le champ de l'entrée : la catégorie grammaticale de l'étymon. Cette indication sert à distinguer des homonymes (cf. l'exemple cité par Büchi 1996, 80 de DESERTUM [adj.] et DESERTUM [s.], FEW 3, 52b). Comme pour la glose, aucune consultation du FEW n'a jamais été évoquée à partir de ce critère. Les structures de l'ouvrage et les besoins des utilisateurs ne justifient donc pas l'identification de cet élément. Toutefois, nous décidons de l'identifier lorsqu'il apparaît, de façon à ne pas invalider la seconde raison donnée ci-dessus comme argument à la non-identification des gloses.

En définitive, la modélisation des articles de la partie étymologisée identifie donc obligatoirement un élément entrée (formalisé par l'élément XML <entry>) contenant, outre du texte éventuel non identifié (correspondant à une glose et éventuellement à une catégorie grammaticale), au moins un étymon-vedette (<etymon>), une langue d'étymon facultative (<lang_etymon>) et une catégorie grammaticale facultative (<gram>). L'étymon-vedette reçoit deux attributs : un attribut *type* ayant la valeur « vedette » et un attribut *lang* ayant pour valeur l'abréviation correspondant à la langue de l'étymon, que cette dernière soit explicitée ou non dans l'entrée de l'article.

La formalisation en schéma XML du champ de l'entrée est la suivante :

```
<xs:complexType name="entry" mixed="true">
    <xs:sequence >
        <xs:element name="etymon" type="etymon"/>
        <xs:choice minOccurs="0" maxOccurs="unbounded">
            <xs:element name="etymon" type="etymon"/>
            <xs:element name="gram" type="gram"/>
            <xs:element name="lang_etymon" type="lang_etymon"/>
        </xs:choice>
    </xs:sequence>
</xs:complexType>
```

Les attributs associés à l'élément <etymon> sont les suivants :

```
<xs:complexType name="etymon" mixed="true">
    <xs:attribute name="type" type="etymon-type" use="required" />
    <xs:attribute name="lang" type="xs:string" use="required" />
</xs:complexType>
```

L'application du modèle à des articles du FEW donne ceci :

> <entry><etymon lang="gr." type="vedette">lakonikos</etymon> <lang_etymon>(gr.) </lang_etymon> lakonisch.</entry> (FEW 5, 133a, LAKONIKOS)
> <entry><etymon lang="lt." type="vedette">ablativus</etymon> ablativ. </entry> (FEW 24, 34ab, ABLATIVUS)

3.5.3.2 L'entrée dans les matériaux d'origine inconnue

Suivant les choix opérés plus haut (→ 3.5.2), nous considérons que le champ de l'entrée dans les volumes 21 à 23 recouvre le concept qui se situe au niveau le plus bas du classement onomasiologique. La modélisation des articles contenant des matériaux d'origine inconnue identifie donc obligatoirement un élément entrée (<entry>) contenant un élément concept (<concept>).

La définition formelle de l'élément <entry> est donc à compléter de façon à intégrer un élément <concept> et à laisser le choix entre un étymon ou un concept (choix qui s'effectue selon l'adresse FEW de l'article) :

```
<xs:complexType name="entry" mixed="true">
    <xs:sequence minOccurs="0" maxOccurs="unbounded">
        <xs:choice>
            <xs:element name="concept" type="concept" />
            <xs:element name="etymon" type="etymon" />
            <xs:element name="gram" type="gram" />
            <xs:element name="lang\_etymon" type="lang\_etymon" />
        </xs:choice>
    </xs:sequence>
</xs:complexType>
```

L'application du modèle dans le texte du FEW donne ceci :

> <art type= "concept" volume="21" pg="148" s="b">
> <entry><concept>colza</concept>.</entry>
> Somme sainse f. „colza“. Tôtes jopière f. „pépinière de colza“.
> [...]
> </art>

Après rétroconversion des volumes 21 à 23, il sera peut-être souhaitable de rétablir l'analyse en structure profonde de Büchi 1996, en identifiant sous chaque concept les différents articles et en intégrant dans ces articles un élément entrée qui sera vide de contenu. Suivant le principe de redescente des informations, la mention du concept pourrait alors être intégrée à l'intérieur de chacun de ces articles, comme attribut de l'élément entrée. Ce dernier comporterait également un attribut *etymon*, qui recevrait la valeur « ? ». Cette valeur pourrait être mise à jour lorsque les matériaux en question

auront été étymologisés. En pratique, cela donnerait ceci pour les articles rangés sous le concept « colza » :

<art volume="21" pg="148" s="b">
<entry etymon = " ?" concept = "colza"/>
Somme sainse f. „colza“.
</art>
<art volume="21" pg="148" s="b">
<entry etymon=" ?" concept = "colza"/>
Tôtes jopière f. „pépinière de colza“.
</art>
[. . .]

Que les concepts soient considérés comme les lemmes d'un article ou comme des éléments macrostructurels de rang plus élevé, leur identification et leur mise en relation directe avec les lexèmes rangés sous eux est nécessaire afin de répondre aux souhaits de consultation transversale portant sur des critères sémantiques. Dans cette optique, il serait intéressant que chaque lexème du FEW, y compris ceux qui se situent dans la partie étymologisée, soit relié à un concept. La présence d'un attribut *concept* au sein de chaque élément <entry> est envisageable pour répondre à ce besoin, mais il existe une solution plus efficace informatiquement parlant, qui consiste à indexer tous les étymons de la partie étymologisée en les associant à un concept. C'est cette solution qui est mise en œuvre à l'ATILF par la constitution d'un index onomasiologique (→ 1.5).

3.5.4 Les renvois internes

Les renvois internes sont les renvois qui, dans un article, mènent à un autre article du dictionnaire. Ils peuvent apparaître à n'importe quel endroit dans l'article et se présentent sous des formes diverses, qui vont de la simple mention d'un étymon-vedette (« S. PAUSARE », FEW 9, 302a, PRAEPONERE 1 b) à une adresse FEW en bonne et due forme, parfois complétée de façon à indiquer une partie précise de l'article (« ici 1, 341b, BESTIA II 1 », FEW 25, 961a, *AURĀTĬCUS n. 6).

Ces renvois jouent un rôle important, car ce sont eux qui assurent les relations entre articles et, donc, la cohésion du discours fewien à un niveau macrostructurel, notamment en réunissant « les éléments d'une super-famille étymologique éclatée dans la macrostructure » (Büchi 1996, 159). Ce rôle est d'autant plus crucial dans une modélisation qui considère chaque article comme une unité autonome. Rappelons que dans le FEW, toute « mise en relation », à quelque niveau où elle se situe, est essentielle pour accéder à l'intégralité de l'analyse fournie. L'identification des renvois internes participe à l'explicitation de la dimension monographique du FEW. Elle participe également à l'optimisation des itinéraires de lecture hypertextuels demandés par les utilisateurs, en facilitant la navigation dans l'ouvrage.

La modélisation du FEW fait donc dégager un élément que nous nommons <renvoi>. Lorsque le renvoi mentionne un étymon, ce dernier est également identifié, ainsi que la langue qui accompagne éventuellement l'étymon. Ces deux informations sont en effet intéressantes dans le cadre de consultations transversales.

La définition en XML d'un élément <renvoi> est la suivante :

```
<xs:complexType name="renvoi" mixed="true">
    <xs:sequence minOccurs="0" maxOccurs="unbounded">
```

```
        <xs:choice>
            <xs:element name="etymon" type="etymon" />
            <xs:element name="lang_etymon" type="lang_etymon" />
            <xs:element name="rpref" type="rpref" />
        </xs:choice>
    </xs:sequence>
</xs:complexType>
```

L'application du modèle sur le texte du FEW donne ceci :

> <renvoi>S. <etymon>pausare</etymon></renvoi> (FEW 9, 302a, PRAEPONERE)

Nous avons pris en compte jusqu'à présent, dans la modélisation, uniquement les renvois à des articles situés dans la partie étymologisée du FEW. Or, des renvois à un article de la partie des Inconnus sont également possibles. Ces renvois internes apparaissent sous la forme d'une flèche suivie de la mention d'un concept :

> Armagnac *garroussin* m. „sainfoin". → vesce. (FEW 21, 148a, « sainfoin »)

Ces renvois sont également identifiés par l'élément <renvoi>. Le concept cité est quant à lui identifié par l'élément <concept>, de la même façon que les concepts situés dans le champ de l'entrée (→ 3.5.2).

Certains articles, qualifiés d'*articles de renvoi*, ne contiennent aucun matériau et sont constitués de ce fait uniquement d'un élément de renvoi (cf. Büchi 1996, 131-133). Ce dernier est identifié de la même façon que les renvois internes situés dans les autres articles, de façon à permettre l'établissement de liens hypertextuels. Les articles de renvoi sont en outre catégorisés explicitement comme tels afin de les distinguer des articles contenant des matériaux. Cette distinction, essentielle pour l'automatisation du processus de rétroconversion, est effectuée par l'ajout à l'élément <art> d'un attribut *type* qui reçoit la valeur « renvoi » s'il s'agit d'un article de renvoi :

```
<xs:attribute name="type" type="article-type"/>
<xs:simpleType name="article-type">
    <xs:restriction base="xs:string">
        <xs:enumeration value="renvoi"/>
    </xs:restriction>
</xs:simpleType>
```

Rappelons que cet attribut est également utilisé pour identifier un article de la partie des Inconnus. Il reçoit alors la valeur « concept » (→ 3.5.2). Les valeurs « renvoi » et « concept » ne sont pas cumulables : la première l'emporte. En effet, la valeur « concept » est utilisée pour identifier un article de la partie des Inconnus contenant des matériaux, ce qui n'est pas le cas des articles de renvoi.

3.6 Modélisation de l'étage inférieur : la cellule lexicale

3.6.1 Identification de la cellule lexicale

L'importance centrale que revêt dans le FEW la cellule lexicale (→ 3.2) suffit à rendre obligatoire son identification, que ce soit dans la partie étymologisée ou dans les ma-

tériaux d'origine inconnue. Nous la formalisons au moyen d'un élément XML nommé <unit>.

3.6.2 Structure de la cellule

Büchi 1996, 116 définit huit molécules composant l'unité minimale de traitement : l'étiquette géolinguistique, le signifiant, la catégorie grammaticale, le signifié, les informations complémentaires, la localisation, la datation et la référence bibliographique. Ces huit molécules sont placées sur le même pied. Or, les quatre premières sont obligatoires (si elles n'apparaissent pas en structure de surface, c'est par ellipse), tandis que les quatre dernières sont facultatives. Cette différence de traitement lexicographique pose question. Si le statut facultatif de la molécule des informations complémentaires se comprend aisément, il n'en va pas de même pour la localisation, la datation et la référence bibliographique : pourquoi ces types d'information seraient-ils facultatifs, alors qu'on connaît l'importance que le FEW leur accorde ?

La réponse se trouve dans le *Beiheft*, ou plus précisément dans l'utilisation implicite que fait le FEW des informations contenues dans cet élément parastructurel du dictionnaire. Dans le FEW en effet, une datation peut renvoyer à une source au même titre que la localisation, et vice-versa :

> [...] au niveau de la structure profonde, la molécule de la datation est toujours implicitement présente, puisque les localisations renvoient à des sources, qui sont, elles, datées : la localisation « Vaux » renvoie à A. Duraffour, *Extrait d'un lexique patois-français du parler de Vaux*, publié en 1923. Tout le matériel n'est donc pas daté, mais tout le matériel est datable à l'aide du *Beiheft*. » (Büchi 1996, 126)

Ces trois molécules, certes différentes quant au contenu, jouent donc un même rôle de renvoi implicite au *Beiheft* qui donne la clé des molécules absentes. Cette fonction de renvoi au *Beiheft* explique qu'elles soient absentes en structure de surface lorsque l'étiquette géolinguistique renferme en elle-même ces informations. Un autre indice de leur parenté étroite est qu'elles apparaissent le plus souvent au sein d'une même parenthèse. Cette analyse structurelle est confirmée par la façon dont le lecteur du FEW perçoit la cellule (→ 2.3.2.3).

Les quatre molécules dites facultatives gagnent donc à être rassemblées dans un même élément, dans lequel elles sont sur le même pied, car aucune d'elles n'est logiquement dominante par rapport aux autres. Nous pouvons dès lors modéliser la cellule lexicale sur deux niveaux. Le premier niveau contient cinq éléments, à savoir les quatre molécules obligatoires et un cinquième élément qui regroupe, à un second niveau, les quatre molécules facultatives (cf. figure 3.3).

Parmi les cinq molécules de premier niveau, le signifiant acquiert facilement dans l'esprit du lecteur une position privilégiée, en représentant le lexème. Ce court-circuitage signifiant → signe linguistique dépasse d'ailleurs largement le cadre du FEW (cf. par exemple Polguère 2008, 36). Le signifiant pourrait donc être considéré comme le noyau de la cellule. Il ne l'est pas en réalité. Le FEW ne traite pas des signifiants, mais des signes complets (signifiant + signifié [+ catégorie grammaticale]). Le fait que signifiant et signifié ne puissent être ellipsés tous les deux dans une même cellule (la mention *id.*

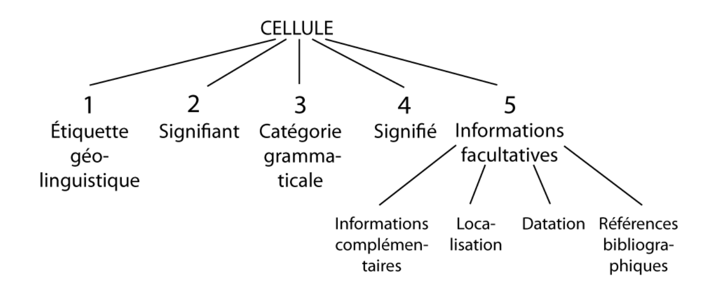


FIGURE 3.3 – Modélisation à deux niveaux de la cellule lexicale

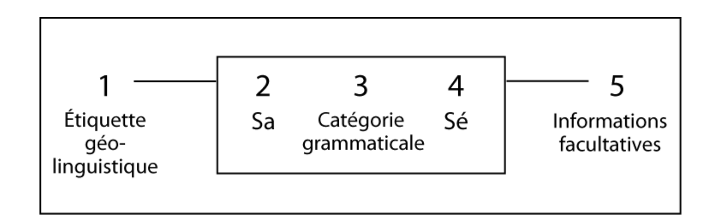


FIGURE 3.4 – Identification d'un noyau dans la cellule lexicale

[pour *idem*] étant alors obligatoirement présente) est, à cet égard, significatif. C'est donc éventuellement le signe complet qui pourrait constituer le noyau de la cellule (cf. figure 3.4).

Ce schéma permet de mettre en évidence les particularités du FEW par rapport à un dictionnaire classique, dans lequel seuls les composant 2, 3 et 4 apparaissent. Nous verrons plus loin que le composant 1 (étiquette géolinguistique) et le composant 5 (élément regroupant les quatre molécules facultatives) fonctionnent en quelque sorte en équipe.

La position en tête de peloton de l'étiquette géolinguistique n'est pas due au hasard. Nous avons déjà dit que cette étiquette était le résultat d'une analyse originale du FEW. Deux cellules possédant les mêmes composants 2, 3 et 4, mais une étiquette géolinguistique différente, ne peuvent être analysées comme contenant le même lexème : ils n'appartiennent pas au même système linguistique (parler, état de langue, ou plus généralement variété linguistique). Nous pouvons dès lors considérer que c'est la somme des quatre premiers composants de la cellule (l'étiquette géolinguistique, le signe composé du signifiant et du signifié, la catégorie grammaticale) qui lui donne son individualité et permet de la différencier des autres cellules. Cette analyse confirme la position sur le même pied des quatre composants obligatoires, qui constituent ensemble le noyau – si l'on peut encore parler de noyau, ce dernier étant assez volumineux – de la cellule lexicale. Ces quatre composants représentent ensemble la cellule lexicale et la concrétisent.

Il nous faut à présent (1) modéliser les quatre composants obligatoires en prenant en compte leur caractère implicite éventuel et (2) comprendre le rôle que joue dans cette cellule le cinquième composant, peut-être plus précieux que ne le laissent supposer son caractère facultatif et sa position à la traîne en fin de peloton.

3.6.3 Modélisation des molécules obligatoires

3.6.3.1 L'étiquette géolinguistique

La reconnaissance des étiquettes géolinguistiques est des plus importantes pour les requêtes demandées par les utilisateurs. L'étiquette géolinguistique constitue en effet une information "originale" du FEW, contrairement aux autres molécules de l'infrastructure (→ 2.2.1). Le choix d'une étiquette géolinguistique est, dans le FEW, moins une donnée fournie par les sources que le résultat d'une analyse, quelquefois très rapide et presque mécanique, quelquefois approfondie, menée par le rédacteur.

Le balisage de l'étiquette géolinguistique permet des consultations transversales sur des critères linguistiques (état de langue, dialecte), mais aussi géographiques et chronologiques, ainsi que bibliographiques. En effet, sous cette étiquette se cachent un grand nombre d'informations implicites accessibles via le *Beiheft*. La création de liens directs entre le *Beiheft* et les étiquettes géolinguistiques du FEW a pour but de faciliter la lecture de l'ouvrage et l'accès à ces informations (→ 2.3.2 ; 2.4.2.2). La première condition à l'établissement de ces liens, outre l'informatisation du *Beiheft*, est évidemment le balisage de l'étiquette elle-même.

La formalisation de cette molécule s'effectue au moyen d'un élément XML que nous nommons <geoling>.

3.6.3.2 Le signifiant

Dans une optique d'informatisation du texte fewien, l'identification des signifiants présente un intérêt évident, puisqu'elle permet de consulter le FEW à partir des lexèmes sans passer par les index partiels. Ce faisant, elle facilite les itinéraires de consultation traditionnels[14].

Parmi les signifiants, il est intéressant de repérer également ceux des locutions, qui font partie des types d'information explicités dans la refonte de la tranche alphabétique B- et sont susceptibles de faire l'objet de consultations ciblées.

La formalisation de la molécule du signifiant s'effectue au moyen d'un élément XML que nous nommons <form>, et qui peut recevoir un attribut *type* indiquant qu'il s'agit d'une locution (type= "locution").

3.6.3.3 La catégorie grammaticale

L'identification de la molécule de la catégorie grammaticale n'offre pas d'intérêt immédiat pour la consultation du FEW en dimension T. En revanche, elle est essentielle en dimension M pour l'opération de découpage des cellules lexicales, ainsi que pour le bon rétablissement de l'implicite infrastructurel (→ 3.6.3.5). Nous décidons donc d'identifier les catégories grammaticales, au moyen d'un élément XML que nous nommons <gram>.

[14]Remarquons qu'il sera nécessaire, lors de la phase d'exploitation, de permettre une recherche non exacte d'un signifiant, en neutralisant notamment les signes diacritiques qui peuvent y apparaître. Ce problème est discuté plus loin (→ 4.2.2 ; 6.5.2).

3.6.3.4 Le signifié

Le balisage des définitions pourrait présenter un intérêt pour des consultations transversales portant sur le sens des lexèmes, en attendant la mise à disposition d'un index onomasiologique du FEW (→ 1.5). Néanmoins, le caractère non rigoureux des définitions du FEW (cf. Büchi 1996, 121-123) les rend très largement impropres à ce type de recherche. Comme pour les catégories grammaticales, leur identification est surtout essentielle en dimension M pour l'opération de découpage des cellules lexicales, ainsi que pour le bon rétablissement de l'implicite infrastructurel (→ 3.6.3.5). Nous décidons de les identifier au moyen d'un élément XML que nous nommons <def>.

3.6.3.5 Traitement de l'implicite infrastructurel

Les quatre éléments <geoling>, <form>, <gram> et <def> que nous venons de définir sont contenus dans l'élément <unit> qui représente la cellule lexicale. Ils sont censés se suivre dans l'ordre, ce qui se formalise comme suit :

```
<xs:complexType name="unit" mixed="true">
    <xs:sequence>
        <xs:element name="geoling" type="geoling" />
        <xs:element name="form" type="form" />
        <xs:element name="gram" type="gram" />
        <xs:element name="def" type="def" />
    </xs:sequence>
</xs:complexType>
```

En pratique, cette succession d'éléments se présente rarement de façon explicite dans chaque cellule. La règle d'économie qui empêche, dans le FEW, de répéter une information déjà donnée précédemment provoque en effet souvent l'ellipse, en structure de surface, d'un ou plusieurs des quatre premiers composants de la cellule (Büchi 1996, 117). Le schéma XML doit donc être modifié de façon à permettre à l'élément <unit> de contenir un, deux, trois ou quatre composants parmi <geoling>, <form>, <gram> et <def> :

```
<xs:complexType name="unit" mixed="true">
    <xs:sequence minOccurs="0" maxOccurs="unbounded">
        <xs:choice>
            <xs:element name="def" type="def" />
            <xs:element name="form" type="form" />
            <xs:element name="geoling" type="geoling" />
            <xs:element name="gram" type="gram" />
        </xs:choice>
    </xs:sequence>
</xs:complexType>
```

Les composants ellipsés ne sont pas pris en compte dans ce schéma. Or, ils doivent absolument être identifiés en dimension T, de façon à assurer l'autonomie de la cellule lexicale et à permettre notamment les itinéraires de consultation demandés par les utilisateurs. Il est également indispensable de les rétablir si l'on veut permettre une mise à jour du FEW qui puisse intégrer les ajouts au sein du texte fewien sans briser

la cohérence syntaxique du discours. En effet, l'ellipse des composants dépend de leur apparition dans les unités précédentes : toute insertion de nouveaux composants sur l'axe syntagmatique est soumise aux mêmes règles et remet en cause l'apparition des composants de même type qui se trouvent à sa suite.

Nous savons que cet implicite peut être rétabli de façon automatique parce qu'il est inférable par la grammaire du FEW (→ 2.6.3.1). La modélisation du FEW explicite donc les composants ellipsés. Cette explicitation ne peut se faire par réinsertion effective de l'information dans le texte lexicographique, puisque nous voulons conserver le reflet du FEW tel qu'il se présente dans son inscription traditionnelle, en modifiant le moins possible sa structure de surface (→ 3.4.3). Le formalisme XML apporte une solution à ce problème en permettant l'utilisation d'éléments vides, c'est-à-dire d'éléments sans contenu textuel. Les composants implicites sont identifiés, là où ils devraient se trouver en structure de surface, au moyen d'un élément XML <imp/>. Cet élément vide de contenu est muni de deux attributs, un attribut *type* ayant pour valeur le type d'information ellipsé (geoling, form, gram ou def) et un attribut *contents* ayant pour valeur le contenu ellipsé :

```
<xs:complexType name="imp" mixed="false">
    <xs:attribute name="type" type="imp-type" use="required"/>
    <xs:attribute name="contents" type="xs:string" use="required"/>
</xs:complexType>
<xs:simpleType name="imp-type">
    <xs:restriction base="xs:string">
        <xs:enumeration value="geoling"/>
        <xs:enumeration value="form"/>
        <xs:enumeration value="gram"/>
        <xs:enumeration value="def"/>
    </xs:restriction>
</xs:simpleType>
```

L'élément <imp> peut dès lors être ajouté aux éléments de la cellule lexicale, de façon à permettre l'identification des molécules ellipsées :

```
<xs:complexType name="unit" mixed="true">
    <xs:sequence minOccurs="0" maxOccurs="unbounded">
        <xs:choice>
            <xs:element name="def" type="def" />
            <xs:element name="form" type="form" />
            <xs:element name="geoling" type="geoling" />
            <xs:element name="gram" type="gram" />
            <xs:element name="imp" type="imp" />
        </xs:choice>
    </xs:sequence>
</xs:complexType>
```

Remarque. Une autre possibilité de formalisation, pertinente au point de vue à la fois de l'analyse linguistique et de l'exploitation informatique des articles rétroconvertis, serait d'utiliser, au lieu d'un élément <imp/>, l'élément XML correspondant à l'information ellipsée. Par exemple, une étiquette géolinguistique ellipsée serait formalisée par un élément vide <geoling/>, contenant en attribut le contenu ellipsé. Cette solution présente l'inconvénient de rendre moins lisible, par un linguiste vérifiant le résultat de

la rétroconversion, la distinction entre composants explicites ou implicites dans le document XML final. Après rétroconversion et vérification de tous les articles du FEW, il est néanmoins envisageable de transformer la balise <imp/> en son correspondant, donné par l'attribut *type* de <imp/>. Le caractère implicite de la molécule serait alors conservé sous forme d'attribut. Par exemple, une balise <imp type="geoling"/> serait modifiée en <geoling imp="true"/>. Le contenu ellipsé (donné par l'attribut *contents* de <imp/>) serait quant à lui, soit conservé tel quel en attribut, soit rétabli explicitement dans le texte du FEW, de la façon suivante :

<geoling imp="true">[contenu]</geoling>

Cette dernière solution ne peut être retenue dans le modèle que nous définissons ici, en raison des principes suivis (→ 3.3). Toutefois, si cette solution s'avérait rentable pour l'exploitation des articles rétroconvertis, elle serait facilement réalisable après la rétroconversion, grâce à une simple transformation XSLT.

3.6.4 Modélisation des molécules facultatives

3.6.4.1 Deux interprétations

Ce que nous avons identifié comme le cinquième composant de la cellule représente sans doute le sous-ensemble le plus complexe à modéliser dans la structure du FEW. Il y a au moins deux façons de l'appréhender.

Une première interprétation suit l'idée selon laquelle le but de ce composant est d'*attester* le lexème. Il est alors perçu comme un *ensemble de références très organisé*, dans lequel chaque molécule est combinée aux autres en structure de surface de façon à remplir collectivement cette fonction d'attestation. C'est l'optique suivie dans la première version de la DTD de la refonte, qui identifiait trois formats de références (→ 1.4.3.7), composés à la base des éléments « Date », « Source », « [source] Secondaire » et « CommentaireSource », avec des combinaisons diverses, dont la « Fourchette », qui représentait une période comprise entre deux attestations.

Dans la pratique, cette DTD très stricte a été abandonnée car elle a montré ses limites : les rédactrices elles-mêmes ont très vite rencontré des situations qui ne rentraient pas dans ces schémas. Nous avons tenté de modéliser ces informations dans cette même optique, mais de façon plus souple, en appelant <attestation> une mention, quelle qu'elle soit, qui indique l'existence du lexème à une période donnée et/ou à un endroit donné (cet endroit pouvant être représenté par un lieu géographique ou par une source, ce qui en dernière analyse revient au même puisqu'à travers le *Beiheft*, une source peut renvoyer à un lieu et vice-versa). L'élément <attestation> est dès lors composé soit d'un élément <localisation> (lieu ou texte), soit d'un élément <datation>, soit d'une combinaison des deux. Une « fourchette » est constituée d'au moins deux attestations qui se suivent. Enfin, la (les) source(s) où ont été trouvées les attestations est/sont citée(s) après l'attestation ou la fourchette. Le composant 5, nommé <references>, pourrait dès lors être défini comme contenant une fourchette ou une attestation, suivie d'une source ou de plusieurs sources, cet ensemble pouvant se répéter

autant de fois que nécessaire. La DTD se présenterait dès lors de la façon suivante :

< !ELEMENT references ((fourchette|attestation), source*)*
< !ELEMENT fourchette (attestation, attestation)>
< !ELEMENT attestation ((localisation ?, datation)|(localisation, datation ?))>
< !ELEMENT localisation (lieu|texte)>

Malgré un résultat qui convient mieux que la DTD de la refonte à la grande variété de formats présents dans le FEW, il nous semble que cette modélisation présente trois défauts, en rapport avec les trois contraintes qui guident la construction du modèle (→ 3.3). Le premier défaut, qui concerne les structures du FEW, est qu'elle est destinée à intégrer l'ensemble des données comprises dans les références : tout extrait de texte ne rentrant pas dans une des cases du modèle défini ci-dessus pose un problème qui remet en question le modèle lui-même. Quand on sait le nombre d'incohérences que contient le FEW, on comprend que cette situation est intolérable : il faudrait que le modèle soit, ou bien tellement complet qu'il prendrait en compte tous les cas particuliers présents dans le FEW, ce qui est pour ainsi dire impossible, ou bien tellement imprécis qu'il conviendrait à tous les cas, ce qui est peu utile. Le deuxième défaut, qui découle du premier, est que son automatisation est sujette à échec, à cause de ces mêmes incohérences. Enfin, le troisième défaut, moins grave, mais qui permet de prendre nos distances avec cette modélisation, est qu'elle va beaucoup trop loin par rapport à ce qui est nécessaire pour répondre aux besoins des utilisateurs.

La seconde façon d'appréhender l'ensemble des quatre molécules facultatives évite ces défauts. Elle consiste à le considérer comme un *ensemble peu organisé, destiné à fournir des informations variables* en fonction des besoins et, plus spécifiquement, des *précisions* par rapport aux informations données par l'étiquette géolinguistique. Il ne s'agirait donc pas obligatoirement d'attester le lexème, mais bien de préciser – si besoin – l'étiquette géolinguistique, soit en l'explicitant, soit en la restreignant. Certes, la plupart du temps, ces précisions consistent en des datations ou des références bibliographiques, ce qui nous ramène à notre première interprétation. Cependant, cette différence de point de vue présente, en ce qui concerne les structures du FEW, deux avantages. Tout d'abord, elle souligne la liberté du rédacteur, qui choisit à chaque moment, parmi les quatre molécules (reliées, rappelons-le, via le *Beiheft*), celle(s)) qui lui semble(nt) la ou les plus pertinente(s) et riche(s) par rapport à l'information qu'il veut transmettre. Ensuite, elle rend hommage à l'analyse faite par Büchi 1996 (123-129), qui distingue les quatre molécules sans chercher à les organiser de façon stricte.

Cette analyse du cinquième composant comme un ensemble de précisions a été confirmée par le directeur actuel du FEW, Jean-Paul Chauveau, dans le cadre de discussions informelles. Il est intéressant de remarquer à ce propos qu'un consensus n'existait pas parmi les rédacteurs anciens et actuels sur les conditions d'apparition de la molécule de la référence bibliographique, ce qui démontre l'absence de discours normatif à ce sujet et l'importance d'une modélisation souple.

Notre cinquième composant peut dès lors être nommé <precisions> et être modélisé comme proposant ci et là des informations intéressantes à identifier, correspondant aux quatre molécules en question, quel que soit leur ordre d'apparition dans le texte fewien. Cette modélisation remplit nos objectifs en ce qui concerne les besoins des utilisateurs et les possibilités d'automatisation, puisque sont identifiés uniquement les types d'information qui sont utiles et reconnaissables de façon automatique.

Cette interprétation très souple présente néanmoins un défaut majeur : elle ne rend

pas compte, ni de l'interrelation qui existe entre les molécules, ni de l'implicite sous-jacent. Par exemple, elle ne permet pas de représenter une fourchette contenant comme *post quem* une date, puis comme *ante quem* une source, qui ensemble définissent une datation. Or, la reconnaissance d'une telle fourchette peut être intéressante, à la fois pour les consultations transversales (une consultation portant sur les lexèmes attestés au 16e siècle bénéficierait de l'identification d'une fourchette telle que 14e – 17e siècle) et pour l'optimisation des itinéraires de lecture (le fait qu'une source cache une date est à expliciter pour faciliter les itinéraires de lecture en dimension M).

Dès lors, nous proposons une modélisation qui, partant de cette interprétation souple de l'élément <precisions>, identifie en son sein, non seulement les types d'information correspondant aux molécules définies par Büchi 1996, mais également les moments pendant lesquels le lexème est attesté, quelles que soient les molécules qui permettent d'exprimer cette information dans le discours. La modélisation de ces différents éléments est décrite ci-après.

3.6.4.2 Identification des localisations

La molécule de la localisation est sémantiquement très proche de celle de l'étiquette géolinguistique. En effet, ces deux molécules peuvent se réaliser au moyen des mêmes termes du lexique fewien (→ 2.4.2.2). Leur nature est donc semblable, même si leur fonction est différente. C'est la position du terme dans la syntaxe fewienne qui détermine s'il joue le rôle d'étiquette géolinguistique ou de localisation. Ce terme sera dorénavant désigné sous l'appellation de *sigle géolinguistique* lorsqu'il est question de sa nature (identique pour l'étiquette géolinguistique et la localisation) et non de sa fonction.

L'identification de tout sigle géolinguistique apparaissant dans le FEW est utile, quelle que soit la fonction qu'occupe ce dernier. Il faut toutefois garder à l'esprit que la molécule de la localisation apparaît en complément à l'étiquette géolinguistique, uniquement lorsque le rédacteur a jugé bon de donner cette information supplémentaire, pour des raisons diverses. En dimension T, la localisation est donc une information moins utile que l'étiquette géolinguistique. En dimension M, il s'agit en revanche d'une information capitale, puisque son apparition indique que l'information géographique donnée par l'étiquette géolinguistique est à nuancer.

En raison de ce qui précède, nous avons décidé que, quelle que soit la fonction qu'ils remplissent dans la syntaxe du discours fewien, les sigles géolinguistiques seront tous identifiés au moyen du même élément <geoling>. Ce choix permet de faciliter l'automatisation du balisage sans nuire à la distinction entre les deux fonctions, puisque la position du terme au sein d'un élément <precisions> suffit pour indiquer qu'il remplit la fonction de localisation[15].

3.6.4.3 Identification des références bibliographiques

Les références bibliographiques du FEW sont variées. On distingue les sources primaires (éditions de texte) et les sources secondaires : glossaires d'édition, lexiques et

[15]Cette assertion est fausse lorsque les précisions contiennent elles-mêmes des cellules lexicales, ce qui arrive de façon tout à fait exceptionnelle seulement. Ce cas particulier est traité plus bas.

dictionnaires, atlas linguistiques. L'identification des sources primaires permet de répondre à des requêtes concernant les auteurs ou les œuvres présents dans le FEW, par exemple en cas de recherches sur les néologismes attestés pour la première fois chez un auteur particulier. Les sources secondaires intéressent également les linguistes et lexicographes. À cet égard, il peut être intéressant de classer les références bibliographiques en quelques grandes catégories, afin de faciliter des requêtes ciblées.

Nous formalisons la molécule de la référence bibliographique au moyen d'un élément nommé <biblio>. Cet élément peut recevoir un attribut *type* indiquant sa catégorie, par exemple « atlas » ou « dictionnaire ». La définition de ces catégories est laissée provisoirement de côté.

Un examen détaillé des références bibliographiques du FEW montre qu'elles sont constituées de sous-éléments variés. Le plus important est ce que nous appellerons le *sigle bibliographique*, qui se présente généralement sous la forme d'une abréviation consignée dans le *Beiheft*. Ce sigle est parfois suivi de précisions diverses, qui diffèrent selon la nature de la source bibliographique : elles peuvent indiquer le volume, la page, le vers, etc. Si l'identification de ces informations n'est pas nécessairement utile dans un premier temps, il est néanmoins essentiel que soit clairement affirmée leur appartenance à la molécule de la référence bibliographique. Cela signifie que l'élément <biblio> doit contenir non seulement le sigle bibliographique, mais également les indications éventuelles qui s'y rapportent, en fonction du format de citation retenu pour la source en question.

Dans cette optique, les références de type « Trév 1771 » ou « Fur 1669 » posent un problème de modélisation. La question consiste à déterminer si la date qui suit le sigle bibliographique fait partie de la molécule de la référence bibliographique ou est à considérer comme une molécule de datation. La DTD de la refonte ne prenait pas en compte cette notation particulière, faisant référence à un dictionnaire et à sa date d'édition, notation pourtant très fréquente dans le FEW. Le *Beiheft* lui-même a mis en évidence le caractère particulier de ce type de référence, en y consacrant une section à part (*4. Chronologisches und sachliches verzeichnis der für das schriftfranzösische benutzten quellen*, section maintenue sous le titre *4. Chronologie des sources lexicographiques utilisées pour le français écrit* dans la 3e édition de 2010). Il faut remarquer que ce type de référence à la lexicographie française pose un problème quant à son statut mixte, à la fois attestation et source secondaire (cf. Büchi 1996, 126).

Nous considérons que les dates d'édition de sources secondaires n'ont pas le même statut que les dates d'attestation provenant de sources primaires. Les premières font, pour nous, partie intégrante du sigle – et donc de la référence bibliographique. En d'autres termes, l'apparition d'une date d'édition à la fin d'un sigle bibliographique ne relève pas d'un choix du rédacteur (choix qui serait guidé par la volonté de préciser l'étiquette géolinguistique), mais d'une règle systématique de rédaction (obligeant le rédacteur à accompagner de sa date d'édition toute citation d'ouvrage lexicographique ayant fait l'objet de plusieurs éditions). C'est la référence bibliographique complète (mention abrégée du dictionnaire + date d'édition) qui est porteuse d'une précision par rapport à l'étiquette géolinguistique. Le fait que cette référence puisse exprimer une information de datation est à mettre sur le compte de l'implicite fewien et du fonctionnement en interrelation des molécules facultatives dans le composant des précisions.

Une référence de type « Trév 1721 » est donc modélisée comme un seul élément <biblio>. Dans le cas de fourchettes du type « Trév 1721-1771 », le second élément

est également identifié comme une molécule de référence bibliographique (<biblio>), avec rétablissement de la partie du sigle ellipsé :

<biblio>Trév 1721</biblio>-<biblio imp="Trév 1771">1771</biblio>

Ces cas particuliers mis à part, le fait que des informations implicites puissent se cacher sous la référence bibliographique, telles que la date ou la caractéristique régionale de la source, est une particularité du FEW qui mérite d'être prise en compte à la fois pour les consultations transversales (dimension T) et pour la lecture du FEW (dimension M). L'accès d'un algorithme à cet implicite de nature pragmatique (→ 2.6.3.4) dépend des modalités d'informatisation du *Beiheft* et n'est pas considéré comme acquis dans la modélisation proposée ici. En revanche, l'établissement d'un lien, par une correspondance terme à terme, entre le sigle bibliographique identifié dans le FEW et le sigle qui y correspond dans le *Beiheft* informatisé apportera une première réponse à ces besoins, en facilitant l'accès de l'utilisateur à ces informations par la constitution d'itinéraires de lecture hypertextuels.

3.6.4.4 Identification des informations complémentaires

La molécule des informations complémentaires peut comporter diverses informations d'ordre diasystématique (cf. Büchi 1996, 123-124). Dans de nombreux cas, il s'agit d'informations données par les sources lexicographiques et retranscrites telles quelles.

L'une de ces informations intéresse particulièrement les utilisateurs : il s'agit du caractère régional d'un lexème. Cette information, notée « rég. », apparaît dans la molécule des informations complémentaires, mais seulement à partir du fascicule 145 (paru en 1984). Les autres informations complémentaires n'ont pas fait l'objet de demandes d'identification de la part des utilisateurs consultés. Elles se révèlent en outre peu systématiques dans leur notation. Enfin, aucune molécule implicite ne se cache derrière la molécule des informations complémentaires.

Suivant le principe de limitation à l'identification des informations utiles, les informations complémentaires ne sont dès lors pas incluses dans la modélisation proposée ici. Leur identification n'est en effet pas essentielle, ni pour les itinéraires de lecture traditionnels, ni pour les nouveaux itinéraires de consultation. Elles restent malgré tout accessibles en dimension T à l'utilisateur intéressé, via une recherche « plein texte », sans que leur balisage soit nécessaire.

3.6.4.5 Identification des datations

Nous considérons comme appartenant à la molécule de datation uniquement les dates explicites qui apparaissent en structure de surface sous des formats divers : dates exactes, siècles ou parties de siècles. Le balisage de ces informations explicites présente un intérêt direct pour les requêtes des utilisateurs portant sur un critère chronologique, ainsi que pour d'autres critères (référence bibliographique notamment) qui seraient implicitement présents dans la molécule de datation. Cette molécule est formalisée par l'élément <date>.

Les datations implicitement données par une autre molécule (référence bibliographique notamment, ou étiquette géolinguistique) ne sont pas rétablies au moyen de cet élément <date>. De même, une fourchette (période comprise entre deux dates) n'est

pas modélisée comme une molécule de datation. En effet, dans de nombreux cas, l'un des éléments de la fourchette n'est pas une date, mais une référence bibliographique. Chacune des deux parties de la fourchette est dès lors considérée comme une molécule à part entière et balisée comme telle. L'identification des fourchettes et des dates implicites est un problème qui ne concerne pas les molécules, mais les relations entre elles.

3.6.4.6 Relations entre molécules : identification des attestations

Les molécules une fois identifiées, il reste à déterminer les relations qui existent entre ces molécules et à rétablir l'implicite sous-jacent. La contrainte d'automatisation qui pèse sur notre modèle ne nous permet pas d'aller très loin dans la résolution de cet implicite. Toutefois, il est possible d'analyser la syntaxe fewienne de façon à repérer les molécules qui, ensemble, fournissent une information chronologique en définissant une période pendant laquelle un lexème est attesté. De cette façon, une fourchette, qu'elle soit constituée (explicitement) de dates ou (implicitement) de sigles bibliographiques, peut être identifiée comme telle.

Dans cet objectif, les différentes molécules qui apparaissent dans les précisions sont regroupées en « propositions ». Tous les types d'information, quels qu'ils soient, qui fonctionnent ensemble appartiennent à une même proposition. Par exemple, une fourchette est considérée comme une proposition regroupant deux molécules. Si cette fourchette est suivie d'un sigle bibliographique indiquant la source lexicographique dont elle provient, ce sigle fait partie de la même proposition.

Nous appelons <attestation> une proposition de ce type. L'élément <attestation> peut contenir une ou plusieurs des trois molécule(s) facultative(s) <geoling>, <date> et <biblio>, ainsi que du texte non identifié (notamment des signes de ponctuation). Il est formalisé de la façon suivante :

```
<xs:complexType name="attestation" mixed="true">
    <xs:sequence minOccurs="0" maxOccurs="unbounded">
        <xs:choice>
            <xs:element name="biblio" type="biblio" />
            <xs:element name="date" type="date" />
            <xs:element name="geoling" type="geoling" />
        </xs:choice>
    </xs:sequence>
</xs:complexType>
```

L'application dans le texte fewien donne ceci :

> (<attestation><biblio>Moz 1812</biblio>-<biblio>DG</biblio></attestation>) (FEW 6/1, 415a, MARTYRIUM III 1 b α a')
> (<attestation>seit <date>ca. 1670</date>, s. <biblio>Trév 1752</biblio> </attestation>) (FEW 4, 623a, ĬNCHOARE II 2 b)

L'identification des propositions d'attestation est essentielle pour la lecture du FEW en dimension M, mais aussi et surtout pour les consultations transversales portant sur des critères chronologiques. Elle doit permettre plus tard, dans un processus d'enrichissement manuel ou semi-automatique, d'associer à chaque proposition les informations implicites qu'elle exprime. Par exemple, dans la première proposition balisée ci-dessus,

il est possible d'expliciter (au moyen d'attributs associés à la balise ouvrante <attestation>, indiquant 1812 comme *terminus post quem* et 1900 comme *terminus ante quem*) la période chronologique exacte que recouvre la proposition. Cette explicitation, si elle est effectuée de façon systématique, permettra à un moteur de recherche de retrouver le lexème lorsque l'utilisateur spécifie une date (par exemple 1880) qui est comprise dans la fourchette.

3.6.4.7 Traitement des molécules obligatoires contenues dans les précisions

Dans les parenthèses de précisions apparaissent parfois des signifiants, des définitions ou des catégories grammaticales :

> Bearn. *esmacherá* „v.a. briser la mâchoire ; v.r. se disloquer la mâchoire“ (dazu *esmacherade* f. „décrochage de mâchoire“). – Nice *desmaisselá* […] (FEW 6/1, 559a, MĀXĬLLA I 1 a).

Lorsque ces types d'information constituent ensemble une cellule lexicale, ils sont regroupés sous un élément <punit>. Une identification par l'élément XML <unit> est exclue : cela compliquerait l'automatisation du traitement et ne serait pas satisfaisant en dimension M, puisque la position d'une cellule au sein d'un élément <precisions> peut être révélateur qu'elle n'est pas à traiter au même rang qu'une cellule lexicale de premier niveau.

Afin que ces types d'information, quelle que soit la raison de leur apparition dans les précisions, soient disponibles pour des consultations transversales, la modélisation du FEW permet l'apparition d'éléments <form>, <def> et <gram> au sein de l'élément des précisions, que ces derniers forment ou non une cellule lexicale. Un attribut *status* ayant pour valeur « contains-unit » indique que les précisions contiennent des molécules qui définissent ensemble une cellule lexicale.

3.6.4.8 Formalisation des précisions

En définitive, le composant que nous avons nommé <precisions> peut être défini de la façon suivante :

```
<xs:complexType name="precisions" mixed="true">
    <xs:sequence minOccurs="0" maxOccurs="unbounded">
        <xs:choice>
            <xs:element name="attestation" type="attestation" />
            <xs:element name="biblio" type="biblio" />
            <xs:element name="date" type="date" />
            <xs:element name="def" type="def" />
            <xs:element name="form" type="form" />
            <xs:element name="geoling" type="geoling" />
            <xs:element name="gram" type="gram" />
            <xs:element name="punit" type="punit" />
        </xs:choice>
    </xs:sequence>
    <xs:attribute name="status" type="precisions-status" use="required"/>
</xs:complexType>
<xs:simpleType name="precisions-status">
    <xs:restriction base="xs:string">
```

```
        <xs:enumeration value="contains-unit" />
    </xs:restriction>
</xs:simpleType>
```

3.6.5 Formalisation XML de la cellule lexicale

En synthèse, la cellule lexicale se formalise comme suit :

```
<xs:complexType name="unit" mixed="true">
    <xs:sequence minOccurs="0" maxOccurs="unbounded">
        <xs:choice>
            <xs:element name="def" type="def" />
            <xs:element name="form" type="form" />
            <xs:element name="geoling" type="geoling" />
            <xs:element name="gram" type="gram" />
            <xs:element name="imp" type="imp" />
            <xs:element name="precisions" type="precisions" />
        </xs:choice>
    </xs:sequence>
</xs:complexType>
```

3.7 Modélisation des étages intermédiaires

Ayant modélisé d'une part l'article et les informations situées aux niveaux supérieurs (macrostructure), d'autre part la cellule lexicale et les informations situées aux niveaux inférieurs (infrastructure), nous devons à présent modéliser les étages intermédiaires, correspondant à la microstructure du FEW. Nous envisageons d'abord le découpage de l'article en quatre champs (→ 3.7.1). Le contenu de la documentation (→ 3.7.2), des notes et du commentaire (→ 3.7.3) est ensuite examiné, ainsi que les relations qui unissent les différents champs (→ 3.7.4). Le traitement de l'implicite microstructurel (→ 3.7.5) clôt l'analyse.

3.7.1 Séparation d'un article en différents champs

Büchi 1996 distingue au sein d'un article quatre champs : l'entrée, la documentation, le commentaire et les notes. Nous avons déjà modélisé le champ de l'entrée (→ 3.5.3). Le champ des notes ne pose aucun problème de modélisation. Il se trouve tout à la fin de l'article et est bien séparé des autres champs. En revanche, la séparation des deux champs de la documentation et du commentaire présente un problème : ils ne sont pas toujours matériellement distincts. Plusieurs articles, essentiellement dans les premiers volumes (1, 2/1 et 3), présentent des cas d'insertion du commentaire au milieu de la documentation, notamment les articles courts, constitués d'un seul paragraphe, mais aussi certains articles plus longs (cf. Büchi 1996, 135).

Si nous cherchons un modèle valable pour tous les articles du FEW, il nous faut accepter que le champ documentaire puisse contenir des parties de commentaire. L'inverse est vrai également : le commentaire peut contenir des cellules lexicales qui constituent en quelque sorte des parties de documentation. Les frontières entre les deux

champs apparaissent donc comme floues, ce qui pose un problème pour l'automatisation du modèle : l'identification automatique de certaines informations nécessite de savoir si elles se trouvent dans une partie de documentation ou dans une partie de commentaire, alors que les frontières d'une partie « parasite » ne sont pas facilement détectables automatiquement.

Afin de permettre aux algorithmes de travailler sur des bases fiables, on peut imaginer un champ « mixte », qui identifierait une partie d'article mêlant documentation et commentaire. Les informations contenues dans ce champ peuvent dès lors être traitées avec davantage de prudence. La question se pose alors de savoir où commence et où se termine un champ mixte. Si nous ne prenons pas comme base de travail une unité précise et reconnaissable de façon automatique, le balisage perd en rigueur et tend vers l'anarchie.

Afin d'éviter ces problèmes, nous choisissons une unité de traitement microstructurelle qui reste stable dans tout le FEW quel que soit l'article : le paragraphe. Il s'agit, certes, d'un élément typographique (et non sémantique), mais il participe à la structure de l'article : dans la documentation, il regroupe des unités de traitement qui partagent un même critère (sémantique, morphologique ou autre). L'intérêt de son balisage est donc réel y compris du point de vue de l'analyse structurelle du FEW. Du point de vue de l'automatisation du traitement, il a l'avantage de constituer un élément discret, facile à cerner.

Cette modélisation conduit tout article à être constitué d'un champ entrée (<entry>) obligatoire, d'un ou plusieurs paragraphes (<p>) et, enfin, d'un champ de notes (<notes>) facultatif. Les paragraphes hors notes sont identifiés comme contenant de la documentation uniquement (<doc>), un commentaire uniquement (<com>) ou un mélange des deux (<mixt>). Dans les paragraphes mixtes, lorsque le discours fewien opère une séparation nette entre la partie documentaire et la partie du commentaire, cette dernière est identifiée (<part-com>).

Les articles contenant un ensemble de paragraphes de documentation suivi d'un ensemble de paragraphes de commentaire pourraient faire l'objet d'un regroupement des paragraphes de documentation en un élement supérieur <doc> et des paragraphes de commentaire en un élement supérieur <com>. Néanmoins, il s'avère que l'exploitation informatique des articles rétroconvertis sera grandement facilitée par un balisage par paragraphe : par conséquent, le regroupement en éléments supérieurs <doc> et <com> ne se justifie pas nécessairement en pratique.

3.7.2 Structuration des éléments de documentation

3.7.2.1 Principes de rédaction

Les matériaux d'un article sont classés et hiérarchisés selon plusieurs critères, afin de les situer exactement dans l'histoire de la famille lexicale qui forme le contenu de l'article. Ces critères sont au nombre de sept : outre le critère de l'étymon immédiat et le critère de la transmission, interviennent les critères phonétique, morphologique, sémantique, géolinguistique et chronologique (cf. Büchi 1996, 105-116).

Contrairement à ce qui se passe dans l'ALW (cf. Boutier 2008), l'ordre des critères n'est pas déterminé. Chaque rédacteur opère pour chaque article une sélection parmi ces critères et les hiérarchise en fonction des spécificités de la famille lexicale traitée

(cf. Büchi 1996, 98-100). La conséquence en est la grande souplesse microstructurelle du FEW, ainsi qu'une très grande variation entre articles, certains étant très fortement structurés et d'autres très peu, voire pas du tout. La variation dans ce domaine est fonction de la matière traitée, du rédacteur et, enfin, de la période de rédaction, le volume des matériaux augmentant avec le temps : les articles de la refonte, notamment, présentent une structuration extraordinairement fine (cf. la préface de Jean-Pierre Chambon au volume 25 ; Chambon 1989b, 330 ; 333), qui semblerait presque excessive en regard de la sobriété des articles du FEW les plus anciens.

3.7.2.2 Marqueurs de structuration en surface

En ce qui concerne la structure de surface, cette organisation microstructurelle se concrétise par trois moyens :

1. le regroupement des matériaux en paragraphes ;
2. la présence de marqueurs alphanumériques qui hiérarchisent les paragraphes ;
3. la présence de signes de ponctuation (point, tiret [semi-]cadratin, point-virgule, virgule) qui délimitent et hiérarchisent différents groupements de lexèmes à l'intérieur d'un paragraphe.

Les cellules lexicales sont donc rassemblées en différents groupes (ensemble de paragraphes, paragraphes, groupes au sein d'un paragraphe) qui sont *hiérarchisés* par des marqueurs alphanumériques (hiérarchisation des paragraphes) et par des signes de ponctuation (hiérarchisation de groupements au sein d'un paragraphe). Ces marqueurs et signes de ponctuation apparaissent dans un ordre déterminé, mais non significatif (cf. Büchi 1996, 100-103).

L'*explicitation* des critères qui ont mené à ces regroupements est réalisée d'une part dans le commentaire, par un texte non structuré qui explique la signification des marqueurs alphanumériques, d'autre part dans la documentation même, par des *marqueurs textuels* tels que *Ablt.* (pour *Ableitungen*), *Redensarten*, etc. qui peuvent apparaître théoriquement à n'importe quel niveau hiérarchique, y compris à l'intérieur d'une section délimitée par un marquage alphanumérique. En pratique, les marqueurs textuels apparaissent souvent là où il n'y a pas de marqueur alphanumérique, de façon à compléter l'explicitation fournie par le commentaire. Ils identifient ainsi, au sein d'un niveau alphanumérique, des groupements de lexèmes qui ne sont pas marqués alphanumériquement, mais qui se détachent malgré tout par la mise en page (paragraphe distinct) ou par la typographie (signes de ponctuation, notamment tiret cadratin).

Dans l'article COMPLETUS par exemple (FEW 2, 982b), le niveau hiérarchique I 1 a est explicité par un marqueur textuel (*Vollständig*) ; au sein de ce niveau, après une première liste de lexèmes, un deuxième groupement de lexèmes est effectué dans un paragraphe distinct. Ce deuxième groupement est explicité par le marqueur textuel *Ablt.* (cf. figure 3.5).

L'article CONIUNCTIVUS (FEW 2, 1053a), parmi beaucoup d'autres, présente sous 1. un exemple de groupements de lexèmes réalisés à l'intérieur d'un même paragraphe par la ponctuation (tiret cadratin). Parmi ces groupements, seul le dernier est explicité au moyen d'un marqueur textuel (*Ablt.*) (cf. figure 3.6).

completus vollständig; vollkommen.

I. 1. a. Vollständig. — Mfr. nfr. *complet* „à quoi il ne manque aucune des parties nécessaires“ (seit ca. 1300, Monstr; Rhlitt 6, 464), saint. St-Seurin *compiet*, Minot *conpiet*, npr. *coumplèt*. — Übertragen. Nfr. *complet* „(pop.) tout à fait ivre“ (seit Flick 1802). — Ablt. Nfr. *se compléter* „achever de s'enivrer“ (pop., seit Littré 1863).

Ablt. — Afr. mfr. *completement* „d'une manière complète“ (13. jh.—1503, TL; Gdf; RF 32, 83), nfr. *complètement* (La Fontaine 1695, dann seit Trév 1752) [1]), sav. *complletament*, lang. *coumpletomen* M. — Nfr. *compléter* „rendre complet“ (seit Trév 1725), saint. *compiéter*, sav. *complletâ*, npr. *coumpleta*, mars. *coumpletar* A. — Nfr. *complètement* „action de mettre au complet“ (seit 1750, text in Fér 1787). — Npr. *coumpletage*, mars. *coumpletàgi* M. — Nfr. *compléteur* „celui qui complète“ (seit Lar 1922, auch adj.).

FIGURE 3.5 – Exemple de groupement de lexèmes effectué dans un paragraphe distinct

coniunctivus verbindend.

1. Mfr. *toille conjoinctive* „(t. d'anat.) membrane muqueuse qui tapisse le globe de l'œil et l'unit aux paupières“ (1372), *conjunctive* f. (1495), mfr. nfr. *conjonctive* (seit Paré), *membrane conjonctive* (Cotgr 1611; wieder seit DG); mfr. intention *conjointive* „sincère, qui vient du cœur“ Froissart. — Nfr. *tissu conjonctif* „tissu qui sépare et unit les autres tissus (t. d'anat.)“ (seit 1863), *conjonctif* (1907—1922), *substance conjonctive* (1863). — Ablt. Nfr. *conjonctivite* „inflammation de la conjonctive“ (seit Raym 1832); *conjonctival* „qui a rapport à la conjonctive“ (seit Besch 1845).

FIGURE 3.6 – Exemple de groupements de lexèmes réalisés à l'intérieur d'un même paragraphe

L'apparition ou non de ces marqueurs textuels varie fortement d'un article à l'autre. Au sein d'un paragraphe, seuls les marqueurs textuels correspondant à des critères morphologiques (tels que les dérivés ou les composés) sont mentionnés de façon quasi systématique.

3.7.2.3 Modélisation

En ce qui concerne les groupements, la modélisation *doit* absolument identifier le découpage de la documentation en paragraphes hiérarchisés par des marqueurs alphanumériques. Cette hiérarchisation des données est en effet un travail important du rédacteur et porte en elle énormément d'informations sur l'histoire de la famille lexicale concernée. Le marquage alphanumérique constitue le squelette de l'article ; c'est ce squelette qui permet aux lexèmes de former ensemble un tout cohérent, chaque lexème étant rattaché à une partie du corps commun par une relation sémantique, morphologique ou autre, selon la signification de chaque marqueur. L'identification du marquage alphanumérique et des groupements qu'il induit est donc essentielle si l'on veut permettre en dimension T des consultations transversales sur des critères morphologiques, sémantiques, de transmission etc. qui, dans la dimension M du FEW, se trouvent à un niveau supérieur à la cellule lexicale.

La modélisation *peut* également prendre en compte la hiérarchisation des matériaux au sein d'un paragraphe, dans la limite de ce qui est possible dans le cadre d'un traitement informatisé. Cette identification est toutefois moins importante dans un premier temps, dans le sens où les relations qu'elle identifie se situent à un niveau d'analyse très fin qui n'est pas concerné actuellement par les attentes des utilisateurs.

Outre l'élément paragraphe, formalisé en XML par l'élément <p>, la modélisation inclut donc les éléments de *groupement* suivants : (1) un élément récursif <struct>, qui regroupe et hiérarchise les paragraphes selon les marqueurs alphanumériques ; (2) un élément récursif <group>, qui regroupe et hiérarchise les cellules au sein d'un paragraphe selon la ponctuation.

La hiérarchisation des paragraphes suivant des marqueurs alphanumériques pourrait se faire de façon à refléter la dimension M du FEW. Dans ce cas, tout niveau hiérarchique serait modélisé par un nouvel élément <struct>, de la façon suivante :

```
<struct>
    <struct>
        <p>1.a. <unit>[. . . ]</unit></p>
        <p><unit>[. . . ]</unit></p>
    </struct>
    <struct>
        <p>b. <unit>[. . . ]</unit></p>
    </struct>
</struct>
```

Nous avons cependant élevé le paragraphe au rang d'unité de traitement. À ce titre, tout comme les deux unités de traitement que sont l'article et la cellule lexicale, il devrait, même sorti de son contexte, conserver les informations qui le concernent. C'est possible si l'élément <struct> n'est pas défini comme un élément supérieur au paragraphe (regroupant plusieurs paragraphes), mais comme un élement lui étant inférieur. Chaque paragraphe comporte dès lors au moins un élément <struct> indiquant le niveau

où ce paragraphe se situe dans la hiérarchie de l'article, de la façon suivante :

<p><struct id="1.a">1.a. <unit>[...]</unit></struct></p>
<p><struct id="1.a"><unit>[...]</unit></struct></p>
<p><struct id="1.b">b. <unit>[...]</unit></struct></p>

La décision d'utiliser un élément <group>, et non l'élément <struct>, pour identifier les groupements de lexèmes au sein de paragraphes permet de simplifier davantage le formalisme. En effet, l'élément <struct> apparaît une seule fois par paragraphe. Il peut dès lors être supprimé en tant qu'élément et être placé comme attribut de l'élément <p> :

<p struct="1.b">b. <unit>[...]</unit></struct></p>

Cette solution est en adéquation totale avec le principe de redescente des informations au niveau des unités de traitement. Elle permet, sans pour autant perdre les informations en dimension M, de rendre chaque paragraphe plus autonome. Cette autonomie est intéressante en dimension T, car les informations microstructurelles se retrouvent par la même occasion déplacées à un endroit connu (le début du paragraphe), accessible (puisque identifié) et plus proche des lexèmes (puisque descendu au niveau du paragraphe). Associer l'information microstructurelle à chaque lexème individuellement, par exemple en ajoutant à chaque cellule lexicale un attribut comprenant sa position dans le plan de l'article, serait trop verbeux. La modélisation proposée, en constituant le paragraphe comme unité de traitement microstructurelle, se veut un juste milieu entre la dimension M et la dimension T du FEW.

L'élément <p> contient dès lors obligatoirement un attribut *struct* qui indique la position du paragraphe dans le plan de l'article.

Afin de permettre les regroupements de lexèmes et notamment l'explicitation du niveau hiérarchique dans cet attribut *struct*, la modélisation doit également identifier les éléments de *marquage* situés en début de paragraphe. Le marquage alphanumérique qui débute un paragraphe est formalisé en XML par l'élément <pnum>. Cet élément reçoit un attribut *id*, qui a pour fonction d'expliciter le niveau hiérarchique exact concerné par le marqueur. En effet, le FEW ne répète pas les niveaux hiérarchiques déjà cités, qui sont donc implicites en structure de surface. Un marqueur « b » situé en début de paragraphe peut équivaloir à « 1 b » ou « 2 b » selon qu'il se trouve cité après « 1 a » ou « 2 a ».

La définition de l'élément <pnum> est la suivante :

```
<xs:complexType name="pnum" mixed="true">
    <xs:attribute name="id" type="xs:string" use="required"/>
</xs:complexType>
```

Le marquage textuel est quant à lui formalisé par un élément <title>. Si ce marquage textuel comporte des sous-éléments intéressants pour des consultations transversales, par exemple la mention d'un étymon ou d'un affixe, ces sous-éléments sont identifiés : un élément <title> peut donc contenir un élément <etymon> ou un élément <affix>, à condition que ces derniers aient été définis dans le modèle (→ 3.7.3). La définition de l'élément <title> est alors la suivante :

```
<xs:complexType name="title" mixed="true">
    <xs:sequence minOccurs="0" maxOccurs="unbounded">
```

```
        <xs:choice>
            <xs:element name="etymon" type="etymon" />
            <xs:element name="affix" type="affix" />
        </xs:choice>
    </xs:sequence>
</xs:complexType>
```

Toujours dans l'optique de consultations transversales, la modélisation identifie également les marqueurs textuels situés à l'intérieur d'un paragraphe et, lorsque c'est possible de façon automatisée, identifie les groupements de lexèmes sur lesquel portent ces marqueurs internes, au moyen de l'élément <group>. Dans le paragraphe 1 de l'article CONIUNCTIVUS par exemple, le groupement de lexèmes marqué par *Ablt.* est identifié comme suit :

> <group><title>Ablt.</title> Nfr. <i>conjonctivite</i> „inflammation de la conjonctive" (seit Raym 1832) ; <i>conjonctival</i> „qui a rapport à la conjonctive" (seit Besch 1845).</group></p>

En définitive, outre l'explicitation de sa position dans le plan de l'article (attribut *struct*), un paragraphe de documentation peut contenir des cellules lexicales (<unit>), un marquage alphanumérique (<pnum>), un marquage textuel (<title>) et des groupements de lexèmes (<group>). Un paragraphe mixte peut contenir également une partie de commentaire (<part_com>).

```
<xs:complexType name="p" mixed="true">
    <xs:sequence minOccurs="0" maxOccurs="unbounded">
        <xs:choice>
            <xs:element name="part_com"     type="part_com" />
            <xs:element name="pnum"         type="pnum" />
            <xs:element name="group"        type="group" />
            <xs:element name="title"        type="title" />
            <xs:element name="unit"         type="unit" />
        </xs:choice>
    </xs:sequence>
</xs:complexType>
```

3.7.3 Structuration des éléments de commentaire et de note

Le champ des notes et le champ du commentaire ont pour particularité de se présenter sous forme de texte suivi non structuré. Les notes fournissent des renseignements complémentaires sur l'un ou l'autre détail (cf. Büchi 1996, 162-164). Le commentaire remplit plusieurs rôles : il explique les matériaux et leur classement, il les situe par rapport aux grandes familles romanes, il fournit une analyse synthétique de la famille lexicale, il critique d'éventuelles étymologies alternatives etc. (cf. Büchi 1996, 136-161).

Les notes et le commentaire contiennent des types d'informations communs qui peuvent être identifiés au sein du texte non structuré dans lequel ils apparaissent. Les étymons, les lexèmes non galloromans, les sigles géolinguistiques, les sigles bibliographiques, les dates et les affixes intéressent les utilisateurs dans le cadre de consultations transversales. L'identification des renvois internes et de certains termes techniques peut

faciliter les itinéraires de lecture. Enfin, l'identification du ou des rédacteur(s) de l'article est intéressante pour contrôler d'éventuelles variantes de rédaction et pour faciliter l'automatisation du modèle (cf. tag-signature).

3.7.3.1 Mention d'étymons

Toute mention dans le FEW d'un étymon est une information intéressante, peut-être davantage encore lorsqu'il ne s'agit pas d'un étymon-vedette. En effet, les principes d'organisation de la nomenclature conduisent à placer en sous-lemmes un certain nombre d'étymons (cf. Büchi 1996, 52).

Les étymons apparaissant dans le commentaire ou les notes sont donc identifiés, dans notre modèle, au moyen du même élément <etymon> que les étymons-vedettes. Il est cependant utile de savoir s'ils remplissent la fonction de vedette, de renvoi ou de sous-lemme. Cette distinction est effectuée selon la position qu'ils occupent dans l'article et est formalisée en XML au moyen d'un attribut type ayant pour valeur « vedette », « renvoi » ou « sous-lemme ».

La langue à laquelle appartient l'étymon est également identifiée, lorsqu'elle est présente en surface, au moyen de l'élément <lang>.

3.7.3.2 Mention de lexèmes non galloromans

Le commentaire est le lieu où la famille lexicale galloromane est replacée dans un contexte plus large (roman). Sont dès lors citées, chaque fois que c'est possible, les formes romanes (les cognats des lexèmes galloromans) ou autres (les emprunts à l'ancêtre commun des langues romanes) qui confirment l'étymologie proposée. Ces mentions sont très intéressantes pour une grande partie des utilisateurs du FEW et sont donc à identifier.

La modélisation des lexèmes non galloromans suit celle qui a été proposée pour les lexèmes galloromans. Les signifiants sont identifiés par l'élément <form>. Les définitions et catégories grammaticales qui l'accompagnent éventuellement sont identifiées respectivement par l'élément <def> et l'élement <gram>. La mention de la langue est quant à elle identifiée par l'élément <lang> (l'élément <geoling> étant réservé aux seules étiquettes géolinguistiques galloromanes).

3.7.3.3 Mention de sigles géolinguistiques

Outre les mentions de langues d'appartenance des étymons et de langues non galloromanes identifiées par l'élément <lang>, il se peut que le commentaire ou les notes contiennent des étiquettes géolinguistiques. Ces dernières sont évidemment intéressantes à identifier, au moyen du même élément <geoling> que dans la documentation.

Si ces étiquettes géolinguistiques précèdent un signifiant et forment avec ce dernier une cellule lexicale, la cellule lexicale tout entière (y compris la définition, la catégorie grammaticale et les précisions éventuelles) est identifiée par l'élément <unit>, de la même façon que dans la documentation.

3.7.3.4 Mention de sigles bibliographiques

Tout sigle bibliographique apparaissant dans les parties de commentaire ou de notes est à identifier, afin de permettre non seulement des consultations transversales portant sur un ouvrage déterminé, mais aussi la création de liens entre ces sigles et leur explicitation fournie par le *Beiheft*.

Les sigles bibliographiques apparaissant hors documentation sont formalisés par l'élément <biblio>, de la même façon que dans les cellules lexicales de la documentation.

3.7.3.5 Mention de dates

Les mentions de datations (dates exactes, siècles ou parties de siècles) apparaissant dans le commentaire ou les notes sont identifiées de la même façon que dans les cellules lexicales de la documentation, par l'élément <date>.

3.7.3.6 Mention d'affixes

Parmi les informations qui intéressent particulièrement les utilisateurs dans le cadre de consultations tranversales se trouvent les mentions de préfixes et de suffixes. En effet, la dimension thesaurus du FEW fait de cet ouvrage le meilleur lieu où étudier la vitalité et l'évolution des préfixes et suffixes latins dans le domaine galloroman. La modélisation du FEW prend donc ce type d'information en considération.

Tous les articles ne mentionnent pas les affixes de façon systématique, loin s'en faut. Les articles de la refonte sont, à cet égard, beaucoup plus explicites que les articles des volumes précédents. Les affixes étymologiques sont généralement cités dans le FEW en petites capitales lorsqu'ils ont donné lieu à une descendance héréditaire (cf. Büchi 1996, 162-164) :

> Wohl -ESTIS + -ELLUS. (FEW 6/1, 8a, MACER n. 8)

La formalisation s'effectue au moyen d'un élément XML nommé <affix>. Un attribut *type* précise s'il s'agit d'un préfixe ou d'un suffixe :

> Wohl <affix type="suffix">-ESTIS</affix> + <affix type="suffix">-ELLUS</affix>. (FEW 6/1, 8a, MACER n. 8)

3.7.3.7 Renvois internes

L'identification des renvois internes, qui peuvent apparaître à n'importe quel endroit d'un article, donc également dans le commentaire, a été justifiée et modélisée plus haut (→ 3.5.4).

3.7.3.8 Termes techniques

L'identification de certains termes techniques allemands (par exemple *entlehnung*) est intéressante dans l'optique d'en fournir une traduction à l'attention des lecteurs non germanophones, en attendant que soit disponible une traduction complète de chaque

commentaire du FEW. Ces termes sont identifiés dans le texte du commentaire par l'élément <term>, qui peut recevoir deux attributs : un attribut *lemme* ayant pour valeur la forme citationnelle (ce qui constituera en quelque sorte un lemme) correspondant à la forme allemande contenue dans le texte et un attribut *fr* ayant pour valeur la traduction française du terme allemand.

3.7.3.9 Signatures d'article ou de parties d'article

La signature équivaut au nom (nom de famille pour les hommes et initiale[s] du prénom et nom de famille pour les femmes) du rédacteur ou des rédacteurs responsable(s) d'un article (cf. Büchi 1996, 160-161). Elle apparaît en principe à la fin du commentaire et en est séparée par un tiret. Plus rarement, une signature indique aussi le responsable d'une partie de l'article, lorsqu'il diffère du rédacteur principal de l'article. Cette partie peut être une note (cf. Büchi 1996, 162), mais aussi un paragraphe de l'article (cf. Büchi 1996, 161, n 218).

Le balisage de la signature d'article présente un évident intérêt intrinsèque. Il permet non seulement de rechercher tous les articles écrits par un rédacteur donné, mais aussi d'avoir un aperçu, à l'échelle de l'ensemble de l'ouvrage, des diverses contributions et de la façon dont les rédacteurs ont pu se répartir la matière. Ces raisons justifient non seulement qu'on balise les signatures présentes en structure de surface, mais également qu'on rétablisse les signatures d'articles là où elles sont implicites.

Le balisage des signatures éventuelles de parties d'article, enfin, se justifie d'une part, pour ce qui est de la connaissance de l'existant, dans un souci d'exhaustivité, et d'autre part, dans la perspective d'une mise à jour permanente du FEW, qui pourrait voir l'insertion de nouveaux paragraphes, pris en charge par des rédacteurs autres que les signataires des articles en question, au sein d'articles déjà rédigés.

L'identification des signatures est réalisée au moyen d'un élément XML nommé <signature>.

3.7.3.10 Formalisation des éléments de commentaire et de notes

Les types d'information définis ci-dessus sont à identifier dans les paragraphes de commentaire, de notes et dans les paragraphes mixtes. La définition des éléments <note>, <com> et <mixt> qui caractérisent un paragraphe doit donc les intégrer. L'élément <note> est par exemple défini comme suit :

```
<xs:complexType name="note" mixed="true">
    <xs:sequence minOccurs="0" maxOccurs="unbounded">
        <xs:choice>
            <xs:element name="affix"   type="affix" />
            <xs:element name="biblio"  type="biblio" />
            <xs:element name="def"     type="def" />
            <xs:element name="etymon"  type="etymon" />
            <xs:element name="form"    type="form" />
            <xs:element name="geoling" type="geoling" />
            <xs:element name="gram"    type="gram" />
            <xs:element name="renvoi"  type="renvoi" />
            <xs:element name="unit"    type="unit" />
```

```
        </xs:choice>
    </xs:sequence>
</xs:complexType>
```

3.7.4 Mise en relation des notes et du commentaire avec la documentation

Une différence importante entre ces deux champs est la façon dont ils dialoguent avec les autres champs, notamment celui de la documentation. Le champ des notes est relié à au champ documentaire et au commentaire par les appels de note, tandis que le commentaire est relié à la documentation par les marqueurs alphanumériques. Appels de note et marqueurs alphanumériques sont à identifier de façon à faciliter la lecture du FEW en dimension M et, en particulier, à permettre les itinéraires de lecture hypertextuels souhaités par la communauté.

3.7.4.1 Mise en relation des notes et appels de note

Afin de permettre la mise en relation des notes et des appels de note, il est tout d'abord nécessaire d'identifier chacune des notes qui se trouve dans le champ des notes et de connaître son numéro. L'élément <note>, identifiant une note au sein du champ des notes (identifié lui-même par l'élément <notes> au pluriel), reçoit dans ce but un attribut *id* ayant pour valeur le numéro de la note, de la façon suivante :

```
<note id="1"> 1)... </note>
```

La définition de l'élément <note> est complétée de façon à intégrer cet attribut :

```
<xs:complexType name="note" mixed="true">
    <xs:sequence minOccurs="0" maxOccurs="unbounded">
        <xs:choice>
            [...]
        </xs:choice>
    </xs:sequence>
    <xs:attribute name="id"   type="xs:positiveInteger" use="required" />
</xs:complexType>
```

Le champ des notes (<notes>) reçoit quant à lui l'information du nombre total de notes qu'il contient, au moyen d'un attribut *count* :

```
<xs:complexType name="notes">
    <xs:sequence minOccurs="1" maxOccurs="unbounded">
        <xs:element name="p"    type="pnote"/>
    </xs:sequence>
    <xs:attribute name="count" type="xs:positiveInteger" use="required"/>
</xs:complexType>
<xs:complexType name="pnote" mixed="true">
    <xs:sequence minOccurs="1" maxOccurs="1">
        <xs:element name="note" type="note" />
    </xs:sequence>
</xs:complexType>
```

Les numéros d'appels de note qui apparaissent dans le texte de l'article sont identifiés au moyen de l'élément <appelnote>. Ils reçoivent également un attribut *id*, qui a pour valeur le numéro de note. Cet attribut permettra la création d'un lien hypertextuel entre un appel de note et la note qui lui correspond. Pour un appel de note contenant le numéro 1, le balisage est le suivant :

> <appelnote id="1"><e>1</e>)</appelnote>

La définition des appels de note est exprimée comme suit :

```
<xs:complexType name="appelnote" mixed="true">
    <xs:attribute name="id" type="xs:positiveInteger" use="required"/>
</xs:complexType>
```

Les appels de note pouvant apparaître à n'importe quel endroit d'un article, l'élément <appelnote> est à inclure dans la définition de tous les autres éléments du modèle, excepté dans l'élément <notes>.

3.7.4.2 Mise en relation du marquage alphanumérique

Les marqueurs alphanumériques présents dans la documentation ont été formalisés par l'élément <pnum> (→ 3.7.2). Les marqueurs présents dans le commentaire et renvoyant à la numérotation du champ documentaire sont formalisés par l'élément <pref> (pour « référence à un paragraphe »). Cet élément reçoit, comme l'élément <pnum>, un attribut *id* explicitant le niveau hiérarchique exact correspondant au marquage cité dans le commentaire :

> Aus COPULATIO in seinen verschiedenen lt. bed. oben <pref id="II 3 a">3 a</pref>, während <pref id="II 3 b">b</pref> wohl durch die schüler aus dem lt. nom. französiert wurde. (FEW 2, 1162a, CŌPŬLA)

La définition de l'élément <pref> est la suivante :

```
<xs:complexType name="pref" mixed="true">
    <xs:attribute name="id" type="xs:string" use="required"/>
</xs:complexType>
```

Outre des marqueurs alphanumériques renvoyant à la documentation du même article, peuvent se trouver dans le commentaire des marqueurs renvoyant à un autre article du dictionnaire. Ces marqueurs sont distingués dans notre modèle par le fait qu'ils se trouvent dans un élément de renvoi :

> <renvoi>ici 1, 341b, <etymon>BESTIA</etymon> II 1</renvoi> (FEW 25, 961a, *AURĀTĬCUS n. 6)

Nous pouvons compléter la modélisation de ces éléments de renvoi par l'identification du marquage alphanumérique qui accompagne parfois l'étymon de renvoi. Cette identification est réalisée par l'élément XML <rpref>, de la façon suivante :

> <renvoi>ici 1, 341b, <etymon>BESTIA</etymon> <rpref>II 1</rpref></renvoi> (FEW 25, 961a, *AURĀTĬCUS n. 6)

L'identification du marquage alphanumérique dans les éléments de renvoi doit permettre la création de liens hypertextuels très précis, menant directement à l'endroit de l'article concerné par le renvoi.

3.7.5 Traitement de l'implicite microstructurel

La modélisation proposée jusqu'à présent pour la structuration de la documentation identifie essentiellement les groupements de cellules. Les critères qui ont présidé à ces regroupements sont explicités uniquement lorsqu'un marqueur textuel est présent, ce qui est assez rare. La mise en relation des marqueurs alphanumériques de la documentation et du commentaire facilite les parcours de lecture, mais n'explicite en rien les critères associés à chaque marqueur : le lecteur doit lire le commentaire pour y chercher une explication éventuelle.

Il serait pourtant intéressant, en dimension T, de disposer de l'explicitation des critères qui ont présidé aux regroupements des lexèmes. En effet, le fait qu'un lexème soit un dérivé ou une locution, caractéristiques pouvant motiver un regroupement, est une information susceptible de faire l'objet de consultations transversales. En dimension M, et en particulier pour les niveaux qui bénéficient d'un marquage alphanumérique, cette explicitation permettrait également de faciliter la compréhension de l'article par la constitution d'une table des matières résumant sa structuration. Cette aide à la lecture est très demandée par les utilisateurs et est d'ailleurs proposée pour certains articles de la refonte de la tranche alphabétique B- (voir par exemple l'article BILANX[16]).

La contrainte d'automatisation qui pèse sur le modèle proposé ici ne nous permet malheureusement pas d'expliciter ces critères de regroupement. En effet, l'explicitation fournie par le commentaire n'est pas assez structurée pour permettre à un automate d'identifier de façon fiable ce type d'information. Cette limitation est une des raisons qui explique le besoin d'un enrichissement du modèle après la rétroconversion du FEW, enrichissement qui ne peut être effectué que de façon manuelle, quel que soit le formalisme qui le permette.

Il est néanmoins possible de répondre partiellement aux attentes des utilisateurs en proposant un plan de l'article qui reprenne, pour chaque niveau de structuration identifié par un marqueur alphanumérique ou par un marqueur textuel, le texte situé au début du paragraphe, au moins jusqu'à la fin de la première cellule lexicale. Il nous semble que de tels extraits de texte pourraient, dans beaucoup d'articles, donner une assez bonne idée au lecteur de ce qui se trouve dans chaque paragraphe. La mise à disposition d'un tel plan de l'article est possible à partir de la modélisation proposée ici. Suivant le principe de non modification du texte original du FEW, le plan ne sera toutefois pas réinjecté dans le texte de l'article. Son affichage à partir du document XML final de l'article rétroconverti sera une fonctionnalité gérée par le logiciel d'exploitation du FEW.

3.8 Application du modèle sur un article du FEW

3.8.1 Résumé des types d'informations identifiés

En résumé, le modèle décrit dans ce chapitre identifie dans le FEW

1. des articles (<art>) qui peuvent être de plusieurs types : article étymologique,

[16]Chauveau Jean-Paul, 2006. BILANX, version provisoire publiée sur le site internet du FEW (`www.atilf.fr/few`), Nancy, ATILF.

article de renvoi, article de concept (ce dernier représentant probablement une solution provisoire, → 3.5.2) ;

2. dans un article : un champ entrée (<entry>), un champ notes facultatif (<notes>) et des paragraphes de documentation (<doc>), de commentaire (<com>) ou mixtes (<mixt>) ;

3. dans l'entrée d'un article étymologique ou de renvoi : un étymon obligatoire (<etymon>) et une langue d'étymon facultative (<lang_etymon>) ;

4. dans l'entrée d'un article de concept : un concept (<concept>) ;

5. dans les paragraphes de documentation : des groupements (<group>), des marqueurs alphanumériques (<pnum>) et textuels (<title>) et des cellules lexicales (<unit>) ;

6. dans une cellule lexicale, cinq informations : l'étiquette géolinguistique (<geoling>), le signifiant (<form>), la catégorie grammaticale (<gram>), le signifié (<def>), les précisions (<precisions>) ;

7. dans les précisions : une ou plusieurs attestations (<attestation>) ;

8. dans les attestations : des références bibliographiques (<biblio>) et des dates (<date>) ;

9. dans les paragraphes de commentaire, une dizaine d'informations ponctuelles : des étymons (<etymon>), des lexèmes et les informations éventuellement associées (<form>, <lang>, <geoling>, <def>, <gram>, <unit>), des sigles bibliographiques (<biblio>), des dates (<date>), des affixes (<affix>), des marqueurs alphanumériques (<pref>), des signatures (<signature>), des termes techniques (<term>) ;

10. dans les paragraphes mixtes : le même contenu que dans les paragraphes de documentation et, éventuellement, une partie de commentaire distincte (<part_com>) contenant le même type d'informations que dans les paragraphes de commentaire ;

11. dans les notes : des paragraphes de note (<note>) contenant un numéro de note et le même type d'informations que dans le commentaire.

12. n'importe où dans l'article : des appels de note (<appelnote>) et des renvois internes (<renvoi>), ces derniers contenant éventuellement des étymons (<etymon>) et des marqueurs alphanumériques (<rpref>).

Ces éléments constituent les types d'information que nous avons jugés pouvoir identifier en vertu des contraintes et des principes qui dirigent la construction du modèle (→ 3.3). Rappelons qu'ils ne constituent pas ensemble un modèle exhaustif qui couvrirait l'ensemble des caractères et des types d'information qui peuvent apparaître dans le texte fewien. En d'autres termes, des parties de texte fewien non identifiées (qu'il s'agisse de mots ou de signes de ponctuation) peuvent apparaître dans le contenu de la majorité des éléments XML de ce modèle.

praeponere voransetzen.

1. a. Apr. *preponer* „résoudre" (Carc. ca. 1200). — Queyr. *prepost* „à propos".

b. S. PAUSARE.

2. Mfr. nfr. *préposition* f. „le fait de placer en avant" (Garb 1487–Pom 1700)[1], „action de manifester (du courage, etc.)" (1531, MirHist 7, 22, Db); „partie du discours, invariable, qui se place entre 2 termes qu'elle lie ensemble en exprimant un rapport de l'un avec l'autre" (seit 1380, Aalma 3239). — Ablt. Nfr. *prépositionnel* adj. „qui a rapport à la préposition" (seit Boiste 1829); *prépositionnellement* adv. „en manière de préposition" (Moz 1842–Besch 1858).

3. a. Mfr. *prépositif* adj. „mis en avant" (1531), nfr. id. Oud 1640.

b. Nfr. *particule prépositive* „préfixe" (Enc 1765–Lar 1874), *prépositif* „qui se met en tête (p. ex. voyelle)" (Boiste 1803–Lar 1875), *locution prépositive* „réunion de plusieurs mots jouant le rôle d'une préposition" (seit Ac 1835). — Nfr. *prépositivement* „à la manière des prépositions" (Besch 1845–Lar 1875).

Lt. PRAEPONERE hat in den rom. sprachen nicht weitergelebt, doch s. PRAEPOSITUS. Das verbum ist vereinzelt im apr. als lehnform belegt (1 a), im fr. in die familie von *poser* aufgenommen worden (b). 2 entlehnt aus *praepositio* „das voranstellen; präposition", 3 a aus *praepositivus* „zum voransetzen bestimmt", b aus dem selben, in der terminologie der grammatiker.

1) Genaue bed. aus den wörterbüchern nicht recht ersichtlich, da diese nur mit *praepositio*, *Vorsetzung*, *a putting before* u. ä. definieren.

FIGURE 3.7 – L'article PRAEPONERE (FEW 9, 302a)

3.8.2 Exemple d'application du modèle

Le modèle que nous venons de construire peut être représenté sous forme de schéma XML, disponible en annexe (→ G.2).

L'application du modèle sur un article du FEW ne met pas en œuvre tous les éléments du modèle, mais uniquement ceux qui sont présents dans l'article. Prenons comme exemple l'article PRAEPONERE (FEW 9, 302a, cf. figure 3.7).

La modélisation de cet article, formalisée en XML de la façon définie ci-dessus, aboutit au résultat suivant :

```
< ?xml version="1.0" encoding="UTF-8" ?>
<few xmlns="http ://www.atilf.fr/few/fsml">
<art book="1" ici="1" id="0" pg="302" s="a" type="doc-com" volume="9">
<entry>
        <b><etymon lang="lt." type="vedette">praeponere</etymon></b>
        voransetzen.
</entry>

<doc>
<p><struct id="1 a" status=" ?">
        <pnum id="1 a">1. a.</pnum>
        <unit>
                <geoling>Apr.</geoling>
```

```
            <form>preponer</form>
            <def>„résoudre“</def>
            <precisions>(
                  <attestation>
                        <geoling>Carc.</geoling> <date>ca. 1200</date>
                  </attestation>)
            </precisions>
      </unit>
      . —
      <unit>
            <geoling>Queyr.</geoling>
            <form><i>prepost</i></form>
            <def>„à propos“</def>
      </unit>.
</struct></p>
</doc>

<doc>
<p><struct id="1 b" status=" ?">
      <pnum id="1 b">b.</pnum>
      <renvoi>S. <sc><etymon type="renvoi">pausare</etymon></sc></renvoi>.
</struct></p>
</doc>

<doc>
<p><struct id="2" status="unknown">
      <pnum id="2">2.</pnum>
      <group status="unknown">
      <unit>
            <geoling>Mfr.</geoling> <geoling status="ok">nfr.</geoling>
            <form><i>préposition</i></form>
            <gram>f.</gram>
            <def>„le fait de placer en<lb/>avant“</def>
            <precisions>(
                  <attestation><biblio type="other">Garb 1487</biblio>
                  –
                  <biblio type="other">Pom 1700</biblio>
                  </attestation>)
            </precisions>
      </unit>
      <appelnote id="1" status="ok"><e>1)</e></appelnote>,
      <unit>
            <imp contents="Mfr." type="geoling"/>
            <imp contents="préposition" type="form"/>
            <imp contents="f." type="gram"/>
            <def>„action de manifester (du courage, etc.)“</def>
            <precisions>(<attestation><date>1531</date>, MirHist 7, 22,
                  <biblio type="other">Db</biblio></attestation>)
            </precisions>
```

```
        </unit> ;
        <unit>
                [. . .]
        </unit>.
</group>—
<group status="Ablt.">
        <title>Ablt.</title>
        <unit>
                [. . .]
        </unit>.
</group>
</struct></p>
</doc>

<doc>
<p><struct id="3 a" status="ok">
        <pnum id="3 a">3. a.</pnum>
        <unit> [. . .] </unit>,
        <unit> [. . .] </unit>.
        </struct></p>
</doc>

<doc>
<p><struct id="3 b" status="unknown">
        <pnum id="3 b">b.</pnum>
        <unit>
                <geoling status="ok">Nfr.</geoling>
                <form type="locution ?"><i>particule prépositive</i></form>
                <def>„préfixe“</def>
                   <precisions>(<attestation><biblio type="other">Enc 1765</bi-
blio>–<lb/>
                <biblio type="other">Lar 1874</biblio></attestation>)</precisions>
        </unit>,
        <unit> [. . .] </unit>
        <unit> [. . .] </unit>.
        —
        <unit> [. . .] </unit>.
</struct></p>
</doc>

<com>
<p><lang>Lt.</lang> <sc><etymon type="vedette">praeponere</etymon></sc> hat in
den
<lang>rom.</lang> sprachen nicht weitergelebt, doch s. <sc>praepositus</sc>.
Das verbum ist vereinzelt im <geoling>apr.</geoling> als lehnform be-
legt (<pref id="1 a">1 a</pref>), im <geoling>fr.</geoling> in die fami-
lie von <form><i>poser</i></form> aufgenommen worden (<pref id="1
b">b</pref>). <pref id="2">2</pref> entlehnt aus <form><i>praepositio</i></form>
<def>„das voranstellen ; präposition“</def>, <pref id="3 a">3 a</pref> aus <form>
```

```xml
<i>praepositivus</i></form> <def>„zum voransetzen bestimmt“</def>, <pref id="3
b">b</pref> aus dem selben, in der terminologie der grammatiker.<signature au-
thor="Wartburg"/></p>
</com>

<notes count="1">
<p><note id="1">1) Genaue bed. aus den wörterbüchern nicht recht ersichtlich, da
diese nur mit <form><i>praepositio</i></form>, <form><i>Vorsetzung, a putting be-
fore</i></form> u. ä. definieren.</note></p>
</notes>

</art>
```

3.9 Conclusion

Ce chapitre avait pour objectif de proposer une modélisation du FEW selon trois paramètres : les structures de l'ouvrage, les besoins des utilisateurs et les possibilités d'automatisation. Chacun des types d'information identifiés dans ce modèle est donc censé appartenir aux structures du FEW, être nécessaire pour permettre les (nouveaux) itinéraires de consultation et de lecture souhaités par les utilisateurs et être reconnaissable de façon automatisée. Les types d'information qui ne répondent pas à un des trois critères de sélection ont été bannis du modèle. Ceux qui sont identifiés constituent un sous-ensemble (en réalité une intersection) parmi tous ceux qui auraient pu être pris en compte dans un modèle exhaustif.

En ce qui concerne les structures du FEW, certaines concessions ont dû être faites à l'automatisation. La décision de regrouper les matériaux rangés sous un concept en un seul article (dénommé article de concept, → 3.5.2) est révélatrice de cet état de fait. D'autres décisions ont été amenées par la nécessité de définir un cadre rigoureux et net pour la formalisation : la constatation d'une limite floue entre le champ documentaire et le champ du commentaire est un problème qui a été résolu par l'invention d'un troisième genre (un champ mixte) et par la décision de ne pas considérer l'article comme une succession de champs, mais comme une succession de paragraphes qui contiennent chacun de la documentation, du commentaire ou un mélange des deux. Il est évident que ces décisions, plus proches de la structure de surface du FEW que de sa structure profonde, constituent des « reculs » par rapport à une modélisation parfaite, reculs qui, dans certains cas, justifieront peut-être une correction manuelle ou semi-automatique à la fin de l'opération de rétroconversion. Néanmoins, la plupart des éléments décrits par Büchi 1996 ont pu être intégrés sans trop de difficultés dans le modèle.

Les besoins des utilisateurs, qui constituent la raison principale du projet d'informatisation, ont été moins malmenés. La plupart des éléments définis dans le modèle sont des informations susceptibles d'être la cible de consultations transversales. D'autres permettent des mises en relation essentielles dans une lecture en dimension M. À cet égard, une limitation importante du modèle est que toute information implicite non inférable grammaticalement ou lexicalement est, par définition, exclue. Nous avons toutefois tenté de dégager le terrain pour des explicitations futures, en allant le

plus loin possible dans l'identification des éléments qui entrent en jeu dans ces processus d'explicitation. La prise en compte de la hiérarchisation de l'article par l'attribut *struct* et par les éléments <pnum>, <pref> et <group> est une tentative dans ce sens, ainsi que le regroupement des molécules facultatives de la cellule lexicale en propositions d'<attestation>. À ce sujet, il est important de souligner que la modélisation proposée ici identifie les groupements uniquement, et non les relations entre les différents éléments. Le fait qu'un lexème (par exemple un déverbal) dérive de l'étymon-vedette de l'article par l'intermédiaire d'un autre lexème (par exemple un verbe) n'est pas explicité dans le modèle, et pour cause : cette relation, totalement implicite, n'est décodable que par un lecteur humain (et de plus un lecteur humain attentif).

La contrainte d'automatisation qui pèse sur notre modèle est donc très forte. Malgré le caractère limité et prudent de la modélisation proposée, il n'est pas certain que tous les éléments décrits dans ce chapitre soient reconnaissables de façon automatique. Seule l'implémentation des algorithmes de rétroconversion décrits dans le chapitre 5 de cette thèse, suivie d'une analyse des résultats (chapitre 6), peut apporter une réponse à cette question.

Le modèle construit se révèle donc un compromis, que nous espérons réaliste, entre ces trois contraintes. Il présente l'intérêt de couvrir la totalité des articles du FEW, courts ou longs, anciens ou nouveaux, du moins dans la partie étymologisée. La majorité des types d'information identifiés par le modèle sont valables également pour les articles des Inconnus : seul le champ de l'entrée nécessitera, eu égard à l'analyse en structure profonde de Büchi 1996, quelques ajustements.

Le discours fewien représente un type très particulier de discours étymologique, qui ne se trouve nulle part ailleurs si ce n'est dans le LEI, qui s'en inspire. C'est pourquoi sa modélisation est unique. Nous avons renoncé à modéliser le FEW selon les standards de balisage XML prônés par la TEI (*Text Encoding Initiative*, `http://www.tei-c.org/index.xml`) : la structure d'un dictionnaire à étymologie intégrante tel que le FEW entrait difficilement dans leurs schémas. La prise en compte *ab principio* de ces standards présentait le risque de perturber l'analyse par des normes extérieures qui n'étaient pas adaptées au FEW. La TEI ayant évolué entre temps vers plus de souplesse, il sera envisageable, après rétroconversion, d'étudier les possibilités de traduction en TEI du balisage XML intégré dans le FEW, de façon par exemple à faciliter la mise en relation de l'ouvrage avec d'autres dictionnaires et bases de données informatisés.

Aucun modèle, quel qu'il soit, n'est parfait. Celui-ci, qui dépend des possibilités d'automatisation et des besoins des utilisateurs autant que de l'analyse des structures internes du FEW, est évidemment provisoire et perfectible. C'est un modèle essentiellement pratique, qui ne constitue en soi qu'une étape dans le processus d'explicitation du raisonnement étymologique propre au FEW et qui pourra par la suite être enrichi en intégrant des formalismes plus avancés. Le modèle proposé dans ce chapitre a le mérite, espérons-nous, de constituer la base formelle sur laquelle pourront venir se greffer les formalismes suivants.

Deuxième partie

Rétroconversion du FEW

Chapitre 4

Architecture du système de rétroconversion

4.1 Introduction

Les résultats que l'on souhaite obtenir sont clairement définis (→ 3.9). Partant d'une version papier (inscription traditionnelle), on souhaite obtenir un document informatique (inscription numérique) muni d'un balisage XML qui le rende exploitable.

Trois étapes interviennent dans le processus d'informatisation, du dictionnaire papier à l'exploitation par les utilisateurs : l'acquisition du texte, la rétroconversion proprement dite et l'exploitation (cf. figure 4.1).

1. La première étape, l'*acquisition* du texte, vise à obtenir, à partir de la version papier du FEW, une inscription numérique comportant le contenu textuel ainsi qu'un balisage minimal (informations typographiques et de mise en page). Une opération de validation clôt cette étape. Elle consiste à vérifier que les documents répondent à un ensemble de normes de base qui les rendent traitables par les algorithmes de rétroconversion.

2. La *rétroconversion* constitue la deuxième étape du processus : il s'agit de transformer les documents par insertion de balises XML, jusqu'à obtention d'un document de sortie où les types d'informations sont identifiés suivant le modèle de balisage XML introduit dans le chapitre 3.

3. Enfin, la troisième étape, qui sort du contexte de cette thèse, est l'*exploitation*

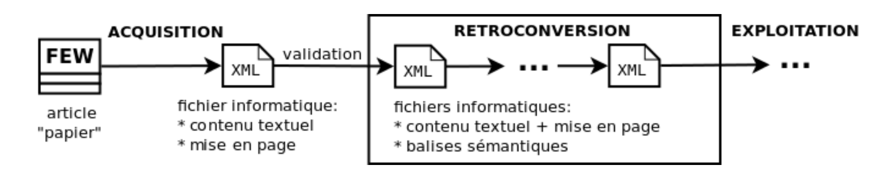


FIGURE 4.1 – Étapes de l'informatisation

des documents rétroconvertis[1].

Le système de rétroconversion proprement dit, objet de cette thèse, concerne donc uniquement la deuxième étape ainsi que, dans une certaine mesure (problématique du codage des caractères et balisage minimal), la première. Trois questions se posent afin de définir un cadre de travail pour la mise au point des algorithmes de rétroconversion :

1. Comment passer du support papier à un support numérique, en tentant de rester aussi indépendant que possible des équipements informatiques ?
2. Quel balisage minimal est-il utile de fournir aux algorithmes de rétroconversion ?
3. Comment concevoir les algorithmes de rétroconversion et les interactions avec les linguistes invoquant le système de rétroconversion ?

Ce chapitre est structuré comme suit : la section 4.2 discute les modalités de passage du support papier au support numérique, la section 4.3 examine comment enrichir le document avec un balisage de base et la section 4.4 décrit le fonctionnement du système de rétroconversion.

4.2 Du papier au support numérique

La première étape consiste à obtenir dans un document informatique le contenu textuel brut (ainsi que des indications de mise en page, → 4.3.2) du document papier. Deux moyens sont envisageables : une numérisation ou une saisie manuelle, qui ont chacune leurs avantages et leurs inconvénients. Quel que soit le procédé choisi, une grande attention doit être portée aux très nombreux caractères spéciaux du FEW.

4.2.1 Saisie par numérisation ou saisie manuelle ?

La numérisation, consistant à scanner le document et à laisser un logiciel d'océrisation reconnaître les caractères, est moins coûteuse que la saisie manuelle dans le cas d'un document volumineux tel que le FEW. En revanche, le résultat d'une océrisation n'est jamais parfait : une inspection manuelle de chaque texte est dès lors obligatoire pour corriger les erreurs. La saisie manuelle, plus coûteuse, conduit à un résultat de meilleure qualité. Une double saisie suivie d'une comparaison semi-automatique entre les deux versions du texte constituerait sans doute la solution idéale, puisque toutes les différences entre les deux versions peuvent être vérifiées et résolues très rapidement, sans relecture complète du texte.

En définitive, le choix entre une numérisation ou une saisie manuelle dépend de deux facteurs : d'abord les moyens financiers, humains et informatiques mis à la disposition du projet, ensuite le niveau de perfection exigé quant au résultat.

En ce qui concerne le FEW, des tests de numérisation effectués à l'ATILF ont permis l'acquisition numérique d'une partie du corpus de test utilisé dans le cadre de cette

[1] Précisons que préalablement à leur exploitation, l'indexation des informations sera nécessaire pour permettre à un moteur de recherche de préparer, dans des temps de réponse acceptables, les résultats aux requêtes linguistiques soumises par les utilisateurs. Cette phase d'organisation des documents rétroconvertis sort également du contexte de cette thèse.

thèse (→ 6.1) avec un pourcentage de reconnaissance avoisinant les 97%. Malgré ces excellents résultats (rendus possibles par l'entraînement du logiciel d'OCR utilisé), de nombreux caractères, dont la reconnaissance était importante pour la rétroconversion (guillemets, caractères en italiques etc.), étaient mal reconnus. Des discussions avec les linguistes et informaticiens allemands Wolfgang Raible (Fribourg-en-Brisgau), Kurt Gärtner et Thomas Burch (Trier), spécialistes de ces questions et intéressés par la numérisation du FEW, ont mené à la conclusion qu'une saisie manuelle était préférable et se révélerait finalement moins coûteuse qu'une numérisation suivie d'une relecture. La saisie manuelle permet par ailleurs de résoudre certains problèmes en permettant de baliser manuellement des informations peu faciles à récupérer lors de l'océrisation. Nous avons donc pris ce mode d'acquisition du texte comme postulat de départ pour l'écriture des algorithmes, avec bon espoir quant à ses possibilités de réalisation, puisqu'un projet de double saisie du FEW a été soumis à une demande de financement en Allemagne[2].

4.2.2 Codage des caractères

Le FEW présente une grande variété de caractères peu communs. Un grand nombre de lexèmes dialectaux sont notés sous leur forme phonique, au moyen de la notation phonétique en usage chez les romanistes et utilisée dans les manuels de phonétique historique du français. Les transcriptions d'étymons provenant de langues orientales, slaves ou autres conduisent à l'utilisation de caractères spéciaux tels que ḏ, ẓ, ğ ou č par exemple. Les étymons grecs et les citations provenant de cette langue se présentent avec leur accentuation et leur diacritisation complète. Enfin, les signes de ponctuation, notamment les guillemets et apostrophes, apparaissent sous des formes particulières (par exemple les guillemets „ “).

Dans une inscription numérique, le codage de tous ces caractères est un problème qui revêt une importance capitale, à la fois pour leur affichage et pour leur traitement par des programmes informatiques. Tout d'abord, il est important de s'assurer que chaque caractère du FEW reçoit le codage adéquat qui lui permettra d'être affiché correctement par la (les) police(s) de caractères utilisée(s), pour autant que ces polices le permettent[3]. Ensuite, il est important de s'assurer, en dehors même des questions d'affichage, que le codage d'un même caractère est uniforme quelle que soit la méthode, le moment ou le contexte de saisie du contenu textuel du dictionnaire. Cette uniformité de codage est nécessaire pour permettre à un algorithme de reconnaître des caractères et de les comparer. Or, à un même glyphe (représentation graphique d'un caractère) peuvent correspondre plusieurs caractères et donc autant de codages : l'océrisation d'extraits du FEW a par exemple conduit à deux codes différents pour le dessin du graphème <n> en italique, pris tantôt, comme il se doit, comme une lettre latine, tantôt, à tort, comme une lettre cyrillique russe (<п>, équivalent de <p>). Si l'œil humain assimile en contexte ces glyphes très similaires, un codage erroné est en revanche désastreux pour un algorithme, deux nombres (codes) différents n'étant évidemment pas équivalents : tout test logique de comparaison de caractères est alors voué à l'échec. Le codage de caractères sur les équipements informatiques a historiquement été une source de difficultés dans

[2]À l'heure où nous écrivons ces lignes, nous n'avons pas encore connaissance des résultats de cette demande de financement soumise à la DFG (Deutsche Forschungsgemeinschaft).

[3]Les polices opérant une sélection plus ou moins large parmi l'ensemble des caractères possibles, il se peut qu'elles ne puissent représenter graphiquement certains caractères.

la mise au point des logiciels, autant pour des raisons d'affichage que de traitement des données.

Heureusement, un standard international maintenant très largement répandu, Unicode (cf. `http://www.unicode.org/`), offre une solution largement supportée par la plupart des équipements informatiques ; celui-ci permet de coder les caractères rencontrés dans la plupart des langues. Le codage Unicode UTF-8, le plus répandu, peut être aisément mis en œuvre pour coder les caractères du FEW, quel que soit le mode de saisie choisi (→ 4.2.1). L'utilisation du codage Unicode résoud une partie des problèmes en fournissant un codage standard (donc utilisé dans la plupart des polices de caractères) pour la plupart des caractères du FEW, tels que les lettres grecques accentuées, les signes de ponctuation particuliers ou encore les signes utilisés dans la transcription d'étymons appartenant à des langues orientales.

Malheureusement, certains signes phonétiques utilisés dans le FEW (un o ouvert long accentué par exemple, ou un « u bouclé ») sont inconnus d'Unicode, de même que certains caractères inédits (par exemple un o barré avec accent circonflexe). Après relevé et analyse, il s'avère que presque *200 caractères du FEW ne sont pas (encore) codables en Unicode*, ce qui constitue pour l'informatisation du FEW un problème remarquable par son ampleur. Ces caractères non standard peuvent recevoir provisoirement un code arbitraire dans une des zones privées de l'espace de codage d'Unicode UTF-8 (zones spécifiquement prévues pour ce type de situation). Ce codage arbitraire et provisoire permet leur traitement par les algorithmes pendant la rétroconversion du FEW. Toutefois, il ne résout ni le problème ultérieur de leur affichage ni le problème, préalable à la rétroconversion, de leur saisie par des opérateurs humains.

Ces deux problèmes d'affichage et de saisie sont bien distincts et requièrent des solutions différentes. Le problème d'affichage nécessite la création d'une police Unicode spécifique au FEW contenant tous les caractères de l'ouvrage avec le codage approprié. La création d'une telle police dépasse le cadre de cette thèse, mais nous nous sommes attelée à la tâche de répertorier l'ensemble des caractères présents dans le FEW et de définir pour chacun (standard et non standard) le codage Unicode adéquat. Le problème de saisie peut quant à lui être résolu plus rapidement grâce au choix du langage XML. La solution consiste à saisir ces caractères sous forme d'entités XML. Ces entités (appelées aussi séquences d'échappement) permettent de nommer un caractère sans le saisir. Il s'agit d'étiquettes aux noms bien choisis et délimitées par les caractères (dits d'échappement) prévus à cet effet que sont l'esperluette et le point-virgule. Nous avons dès lors défini, pour chaque caractère non standard du FEW, des noms d'étiquettes qui sont intuitivement compréhensibles pour tout linguiste ayant un minimum de culture en phonétique. L'entité *&few-o-ouvert-bref;* est par exemple utilisée pour saisir le caractère *ǫ̆* qui n'est pas disponible en Unicode. Les caractères Unicode peuvent également être représentés sous forme d'entité : dans ce cas, le nom de l'étiquette est son code Unicode. L'entité *ã* représente par exemple le graphème *ã* (a tilde) dont le code Unicode est U+00E3.

Le résultat de nos travaux est une table, que nous espérons la plus exhaustive possible, de tous les caractères apparaissant dans le FEW (→ annexes), accompagnés de diverses informations à leur propos. Cette table est utilisable à différents moments de l'informatisation du FEW : d'abord comme guide lors de la saisie, ensuite comme base de données lors de la rétroconversion, enfin comme protocole pour la création d'une police FEW spécifique.

Glyph	FFML & keyword encoding		FSML encoding	Flattening	Delimiter	Greek	Phonetic
	XML escape sequence	UTF-8	Unicode	ASCII	Flag	Flag	Flag
ă *ă*	ă		U+0103	a	.	n	y
ā *ā*	ā		U+0101	a	.	n	y
á *á*	á		U+00E1	a	.	n	y
Á	Á		U+00C1	A	.	n	y
à		à	U+00E0	a	.	n	y
â		â	U+00E2	a	.	n	y
ä *ä*		ä *ä*	U+00E4	a	.	n	y
Ä	Ä		U+00C4	A	.	n	y
å *å*	å		U+00E5	a	.	n	y
ã	ã		U+00E3	a	.	n	y
ắ	ắ		U+1EAF	a	.	n	y
ǟ	ǟ		U+01DF	a	.	n	y
ā́	&few-a-long-accent;		U+F400	a	.	n	y

FIGURE 4.2 – Table de caractères du FEW

La table de caractères inclut pour chaque caractère huit informations (cf. figure 4.2). Les quatre premières sont les suivantes :

– un ensemble de glyphes (si disponibles) correspondant au caractère ;

– l'entité XML (le cas échéant, pour les caractères rares) à utiliser lors de la saisie ;

– le caractère codé en Unicode UTF-8 (le cas échéant, pour les caractères standards) à saisir tel quel lors de la saisie ;

– une chaîne de caractères exprimant la valeur hexadécimale du code Unicode UTF-8 (soit standard, soit assigné dans la zone privée d'Unicode arbitrairement choisie pour la rétroconversion du FEW et commençant à 0xF400, conventionnellement notée U+F400), à utiliser pour la rétroconversion et pour la création d'une police Unicode.

Les quatre dernières informations servent uniquement lors du processus de rétroconversion. Elles sont utilisées par certaines opérations de détection de mot-clé intervenant dans les algorithmes (→ 5.2.2.3). Ces informations sont les suivantes :

– le caractère de la version *aplatie* du caractère. L'*aplatissement* d'un caractère accentué (ou muni de tout autre signe diacritique) consiste à le ramener au caractère standard (c'est-à-dire ASCII) correspondant. Par exemple, <é> est aplati en <e>. La version aplatie sert notamment à des opérations de classement alphabétique ou à des recherches ne prenant pas en compte les diacritiques ;

– un drapeau booléen[4] indiquant si le caractère est un « délimiteur de mots » au sens du FEW. Nous appelons *délimiteur de mot* tout caractère qui n'appartient pas à un mot (signes de ponctuation, caractères d'espacement, chiffres etc.) ;

– un drapeau booléen[5] indiquant si le caractère est une lettre grecque ;

– un drapeau booléen indiquant si le caractère est un caractère phonétique.

[4]Le caractère « . » signifie que la propriété est fausse ; le caractère « | » signifie que la propriété est vraie.

[5]Le caractère « n » signifie que la propriété est fausse ; le caractère « y » signifie que la propriété est vraie.

Certains des caractères non Unicode présents dans le FEW auraient pu être codés au moyen d'une combinaison de caractères Unicode (lettre + signes diacritiques). Par exemple, le o fermé bref accentué (*&few-o-ferme-bref-accent;*) pouvait recevoir le codage 1ECD 0306 0301, combinant le o fermé, le signe diacritique bref et l'accent. La raison pour laquelle cette solution de codage n'a pas été retenue est double. En ce qui concerne l'affichage tout d'abord, ces codes combinés conduisent à des résultats très variables et parfois désastreux en fonction des polices et des logiciels utilisés. Ensuite, au niveau du traitement informatique, ils présentent le désavantage d'être représentés par plus d'un caractère, ce qui complique leur traitement par les outils de développement logiciel et conduit à une pénalité de performance. Le fait que chaque glyphe du FEW soit représenté, lors de la rétroconversion, par un seul caractère Unicode permet d'utiliser des outils de programmation standard.

4.3 Du document numérique brut à un document XML avec balisage de base

Les problèmes d'acquisition du texte brut étant réglés, trois autres questions se posent avant de développer le logiciel de rétroconversion. La première question concerne la quantité de données à rétroconvertir de manière atomique, en une seule fois. La deuxième question a trait à la récupération des informations typographiques et de mise en page de la version papier. Ces deux questions amènent à prendre des décisions qui influencent la construction du logiciel de rétroconversion. La troisième question consiste dès lors à déterminer comment vérifier qu'un document soumis au logiciel de rétroconversion suit les décisions prises, condition nécessaire pour être traitable par ce dernier.

4.3.1 Découpage du dictionnaire en unités atomiques de rétroconversion

Le choix de la quantité de données à rétroconvertir en une seule fois n'est pas une question triviale. Certains projets d'informatisation similaires au nôtre[6] traitent l'ensemble d'un dictionnaire en plusieurs phases successives et synchronisées, ce qui signifie que tous les articles sont traités en même temps. Cette méthode permet de repérer et de corriger très rapidement, pour l'ensemble des articles du dictionnaire, des erreurs fréquentes, avant de passer à la suite du traitement. L'inconvénient de cette méthode est qu'elle nécessite des moyens financiers, humains et logiciels importants. Le contenu textuel de l'ensemble du dictionnaire (acquisition du texte brut) doit être disponible avant d'obtenir le premier résultat. La taille des données à traiter pose elle aussi, dans le cas de dictionnaire volumineux, des difficultés qui nécessitent d'être prises en compte par l'implémentation logicielle du système de rétroconversion.

Dans le cas du FEW, les moyens financiers et humains disponibles pour le projet n'étaient pas encore définis au moment de commencer cette thèse et ne le sont toujours pas. Une des craintes qui rendait utopique l'informatisation de cet ouvrage volumineux et justifiait le besoin d'une étude préalable était précisément le coût de cette opération.

[6] Voir par exemple les projets en cours à l'*Instituut voor Nederlandse Lexicologie* à Leiden (`http://www.inl.nl/`).

Or, de nombreux linguistes espèrent, notamment dans le cadre de projets particuliers[7], que des articles du FEW rétroconvertis seront assez rapidement disponibles. Dès lors, le défi consiste pour nous à imaginer un système de rétroconversion qui, même en l'absence de financement, permette de mettre assez vite à la disposition de la communauté scientifique un échantillon d'articles intéressants. Dans cette optique, une rétroconversion du FEW en une seule fois a été écartée.

La question s'est alors posée de savoir si le FEW devait être rétroconverti par volume, par fascicule, par article, par section d'article, par paragraphe, ... Le choix de l'unité atomique de rétroconversion s'est porté sur l'article pour les raisons suivantes. D'une part, la dimension monographique du FEW ($\rightarrow$ 2.5) interdit de rétroconvertir séparément des éléments de niveau inférieur à l'article. L'article forme en effet un tout cohérent dont le contenu n'est pas dissociable. On confirmera plus loin que cette propriété a un sens du point de vue algorithmique, puisque la détection de certains types d'informations dépend d'informations présentes dans d'autres parties de l'article (détection, par exemple, d'informations du champ du commentaire effectuées sur la base d'informations du champ documentaire, $\rightarrow$ 5.4.12). Bien plus, un découpage du FEW article par article présente, par rapport à un découpage en unités supérieures (volume, fascicule ou section macrostructurelle), au moins quatre avantages :

– Un traitement par article permet de rétroconvertir le FEW petit à petit, article par article, et non en un seul bloc ; les articles rétroconvertis peuvent être rendus disponibles pour la communauté scientifique et pour le public sans attendre que la totalité du dictionnaire ait été traitée.

– Un traitement par article permet de ne pas procéder de manière linéaire (dans l'ordre d'apparition des articles dans le dictionnaire), mais par priorité, en sélectionnant des articles ou des ensembles d'articles correspondant aux besoins d'un projet de recherche donné ; il permet dès lors une planification plus souple du processus de rétroconversion et le choix des priorités en fonction de critères scientifiques (domaines étymologiques par exemple) davantage que de critères physiques (volume par exemple).

– Dans le cadre d'interventions manuelles, un traitement par article, par son systématisme, rend plus facile la détection des erreurs et leur correction.

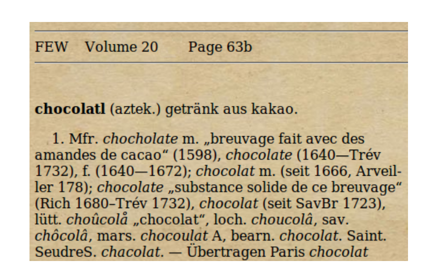
– Dans le cadre de cette thèse, de plus, un traitement par article permet la vérification de l'informatisation sur un premier ensemble d'articles ($\rightarrow$ 6).

Un traitement séparé de chaque article du FEW a dès lors été choisi. Le système de rétroconversion n'accepte un fichier contenant plusieurs articles (par exemple un volume complet du FEW) qu'à la condition que chaque article y ait été préalablement distingué, lors de la saisie, au moyen d'un balisage spécifique (<art>...</art>).

4.3.2 Introduction d'un balisage typographique

Lors de l'acquisition du texte, que ce soit par numérisation ou par saisie manuelle, il est important de récupérer certaines informations typographiques et de mise en page. Un

[7]Par exemple le projet ANR DETCOL (Colombat/Pelfrêne/Buchi 2006, `http://ctlf.ens-lsh.fr/documents/ct_projet_detcol.pdf`) ; voir aussi Alletsgruber à paraître.



FEW Volume 20 Page 63b

chocolatl (aztek.) getränk aus kakao.

1. Mfr. *chocholate* m. „breuvage fait avec des amandes de cacao“ (1598), *chocolate* (1640—Trév 1732), f. (1640—1672); *chocolat* m. (seit 1666, Arveiller 178); *chocolate* „substance solide de ce breuvage“ (Rich 1680–Trév 1732), *chocolat* (seit SavBr 1723), lütt. *choûcolå* „chocolat“, loch. *choucolâ*, sav. *chôcolâ*, mars. *chocoulat* A, bearn. *chocolat*. Saint. SeudreS. *chacolat*. — Übertragen Paris *chocolat*

FIGURE 4.3 – Exemple de visualisation d'un article du FEW

balisage classique des grasses ( ), exposants (<e> </e>), italiques (<i> </i>), petites capitales (<sc> </sc>), marqueurs de fin de colonne (<col/>) et de fin de ligne (<lb/>) est tout à fait possible et souhaité.

Par ailleurs, le choix d'une saisie manuelle comme mode d'acquisition du texte (→ 4.2.1) permet d'extraire, par inspection visuelle, un certain nombre d'informations non textuelles très utiles. En particulier, la saisie du FEW permet la récupération de métadonnées indiquant le positionnement de l'article dans le dictionnaire (volume, tome, numéro de page et de colonne, numéro d'ordre de l'article dans la colonne : <art volume="3" book="1" ici="1">, <col pg="451" s="b"/>) ainsi que des métadonnées structurelles : en-tête (<h>...</h>), paragraphes (<p>...</p>) et zone de notes (<notes>...</notes>).

Les noms de ces balises ne suivent pas nécessairement les spécifications de la TEI (*Text Encoding Initiative*, cf. `http://www.tei-c.org/index.xml`). Ils ont été choisis avant tout pour leur brièveté, afin de réduire le coût de leur saisie, tout en conservant une grande interprétabilité.

L'extraction de ces informations typographiques, positionnelles et structurelles est nécessaire pour le bon fonctionnement des algorithmes de rétroconversion : par exemple, la mise en italiques constitue un critère essentiel pour la détection des formes (signifiants) dans les cellules lexicales (→ 5.4.17). Ces informations sont d'autre part importantes d'un point de vue plus matériel, puisqu'elles permettent de composer, de manière basique, une visualisation sur écran de tout article (cf. figure 4.3).

En définitive, chaque document soumis au système de rétroconversion correspond donc à un article du FEW, contenant des indications typographiques et de mise en page importantes (notamment numéros de page, volume, éventuellement tome). Un ensemble de consignes et conseils de saisie, ainsi que les spécifications du balisage XML de base ci-introduit (dénommé *FEW Font-style Markup Language* ou FFML), est proposé dans un document à destination de la communauté internationale et repris en annexe (→ D).

4.3.3 Validation des documents XML avec balisage de base

Une validation de tout document textuel contenant un balisage XML de base (document au format FFML) est indispensable, de manière à vérifier qu'il est conforme au format d'entrée attendu par le système de rétroconversion. Il est en effet nécessaire pour construire un système d'émettre un minimum d'hypothèses sur les données à traiter. Toute spécification définie et garantie évite autant de problèmes ultérieurs dus à des

données non reconnues.

L'élaboration d'une table des caractères du FEW (→ 4.2.2) permet en outre de vérifier qu'aucun caractère non licite, en particulier les caractères non imprimables et non ASCII qui ne sont pas détectables par inspection visuelle (par exemple les caractères d'espacement non standard[8]), n'est introduit dans le document.

L'implémentation logicielle du système de rétroconversion valide[9] dès lors, avant de le traiter, tout fichier XML contenant un article saisi. Les problèmes éventuels sont signalés. Un fichier XML ne peut donc être rétroconverti que si son contenu (un article du FEW) est codé au format Unicode UTF-8 selon la table de caractères licites du FEW et s'il comporte uniquement le balisage de base (typographique et de mise en page) autorisé, c'est-à-dire respectant le schéma défini pour FFML.

4.4 Du document XML avec balisage de base au document final

4.4.1 Architecture du système de rétroconversion

La rétroconversion proprement dite d'un article consiste à insérer un balisage XML dans un document répondant aux spécifications définies ci-dessus (→ 4.3). Le balisage, conforme à la modélisation définie au chapitre 3, identifie de nombreux types d'information différents. Les problématiques principales du volet informatique de cette thèse sont donc (1) de trouver des indicateurs permettant la reconnaissance automatique de chaque type d'information et (2) d'écrire un algorithme capable d'utiliser ces indicateurs pour insérer toutes les balises aux bons endroits dans le texte d'un article.

Il est important de noter que dès le début de notre recherche, l'objectif premier qu'il nous était demandé de remplir n'était pas l'écriture d'un algorithme optimal, mais le balisage d'un maximum d'informations et, surtout, l'exactitude de ce balisage malgré les nombreuses incohérences structurelles du FEW. Nous avons dès lors décidé d'examiner chaque type d'information l'un après l'autre afin de déterminer, pour chacun, s'il était possible de le reconnaître de façon automatique. Cet examen a conduit rapidement à la constatation que plusieurs types d'information n'étaient pas reconnaissables sans l'identification préalable d'autres types d'information. Dès lors, dans cette optique d'exactitude du balisage, nous en sommes logiquement venue à un système de rétroconversion où chaque type d'information est balisé dans le texte par un algorithme spécifique, les différents algorithmes se succédant selon un ordre bien défini.

Environ quarante algorithmes raffinent successivement le balisage du document textuel représentant l'article à rétroconvertir. Le processus part d'un fichier XML au format FFML (→ 4.3) et ajoute des balises pour obtenir des fichiers XML au format FSML (pour *FEW Semantic-style Markup Language*), offrant une structuration de l'article de plus en plus complète. Ce système de rétroconversion crée autant de fichiers intermédiaires qu'il existe d'algorithmes dans la séquence de rétroconversion. À chaque article est donc associé un ensemble de *N* fichiers XML : un fichier XML au format FFML et plusieurs dizaines de fichiers au format FSML (cf. figure 4.4).

[8] Voir par exemple le caractère U+2028 (`http://www.fileformat.info/info/unicode/char/2028/index.htm`).

[9] Validation SAX avec l'implémentation Java du parser XML Apache Xerces.

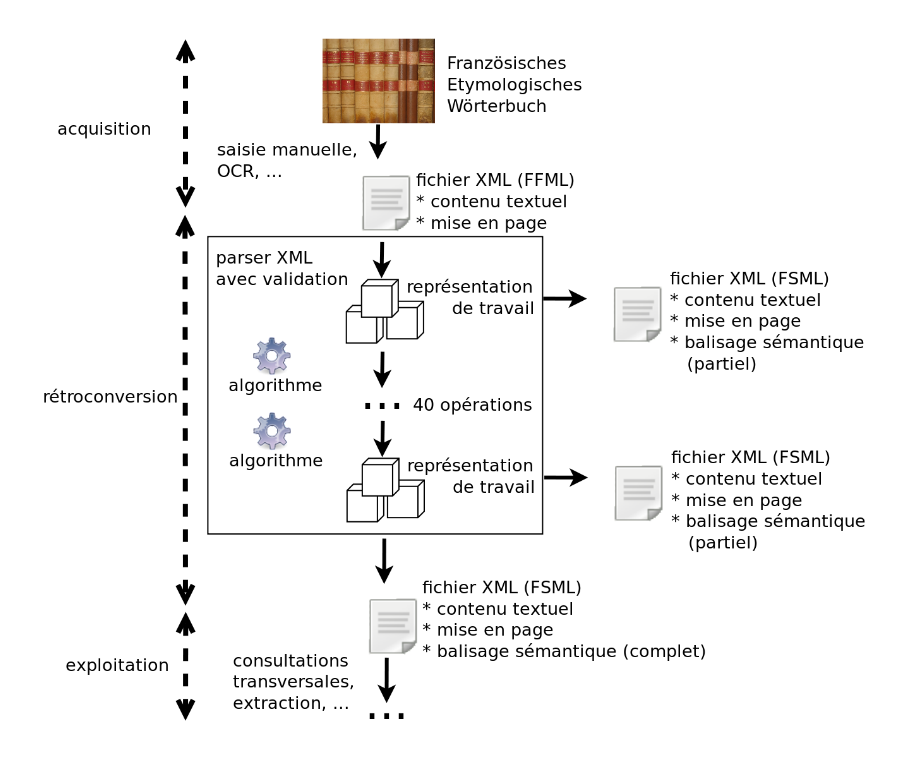


FIGURE 4.4 – Architecture du système de rétroconversion

De manière interne, dans la mémoire de l'ordinateur, le système de rétroconversion opère non sur des fichiers XML, mais sur une représentation de travail exposée dans la section suivante (→ 4.4.2). Après l'exécution de chaque algorithme, le système de rétroconversion génère le fichier XML au format FSML correspondant à l'état courant de la représentation de travail. Tout comme le fichier XML initial, ces fichiers XML sont validés au fur et à mesure, de manière à garantir qu'ils respectent bien le format FSML attendu.

4.4.2 Représentation du texte traité par les algorithmes

L'écriture des algorithmes nécessite une définition précise de l'objet auquel ils seront appliqués. Une question se pose quant à la façon dont est représenté le texte du FEW traité par les algorithmes et, plus particulièrement, les balises qui y sont insérées. Ces dernières doivent-elles être considérées comme des caractères normaux, au même titre que les autres caractères du FEW, ou comme des éléments particuliers hors texte ?

Nous avons défini, pour le processus de rétroconversion du FEW, trois représentations possibles du texte. Ce dernier peut être traité comme une chaîne de caractères, comme une séquence de *chunks* ou comme une chaîne de caractères virtuelle.

4.4.2.1 Première représentation : chaîne de caractères

Un document répondant au format FFML (→ 4.3) comporte une série de caractères codés en Unicode. Les balises XML qui viennent s'insérer dans ce document sont composées également d'une séquence de caractères Unicode, introduite par un chevron ouvrant "<" et terminée par un chevron fermant ">". Le texte traité par les algorithmes se résume donc en une longue chaîne de caractères.

Ce modèle linéaire implique une différence par rapport à la représentation visuelle qui est celle de la version imprimée du FEW : la dimension spatiale ne peut être utilisée par les algorithmes. Ce qui se trouve *à gauche* (ou *à droite*) d'un caractère sera donc tout ce qui le précède (ou le suit) dans la chaîne de caractères, quelle que soit la disposition spatiale du texte dans sa version imprimée[10].

4.4.2.2 Deuxième représentation : séquence de *chunks*

Le processus de rétroconversion a ceci de particulier que le texte du FEW est balisé au fur et à mesure de l'application des algorithmes. Le document XML est en construction : sa structuration est, avant l'application du dernier algorithme de la séquence, toujours imparfaite. En outre, l'ordre d'insertion des balises ne suit pas nécessairement des rapports logiques d'inclusion (par exemple, du plus haut niveau de la structure jusqu'au plus bas, ou l'inverse). Ces particularités expliquent que la structuration du document réalisée par les balises XML soit, pour nous, moins intéressante que la linéarité du texte : les algorithmes de rétroconversion traitent avant tout le texte lui-même et non sa structure encore imparfaitement balisée. Toutefois, les balises insérées par les algorithmes précédents peuvent être utiles, dans un algorithme, pour vérifier certaines propriétés du texte. Nous avons besoin de pouvoir détecter la présence de balises à certains endroits du texte et d'accéder aux attributs qui y sont associés, sans que la structure d'arbre XML soit disponible.

Il est donc nécessaire, tout en conservant une représentation séquentielle de la chaîne de caractères, de pouvoir accéder aux balises en les considérant comme des informations particulières de type XML, dissociées du texte fewien.

À partir de la représentation du texte comme chaîne de caractères, nous construisons une deuxième représentation dont l'objectif est d'opérer, dans la chaîne de caractères, une distinction entre le texte du FEW et les balises XML qui y sont insérées. La chaîne de caractères est subdivisée en sections (*chunks*) qui diffèrent par leur type : une section de balisage (*tag chunk*) est une balise XML, tandis qu'une section de texte (*text chunk*) est une séquence de caractères provenant du texte fewien. Dans le champ de l'entrée de l'article COMPLETUS (FEW 2, 982b), par exemple, le résultat de la rétroconversion produit 6 *tag chunks* et 2 *text chunks*, dont la position dans la séquence de *chunks* est connue. Cette représentation peut être schématisée comme suit, les barres verticales indiquant une séparation entre les différentes sections :

> |<entry>||<etymon>|completus|</etymon>|| vollständig ; vollkommen.|</entry>|

[10]Lors de l'écriture des algorithmes, nous utiliserons donc les seules notions de "contexte gauche" et de "contexte droit" avec cette signification linéaire.

4.4.2.3 Troisième représentation : chaînes virtuelles

La représentation décrite ci-dessus a pour avantage d'opérer une distinction entre les balises et les sections de texte. Les inconvénients sont que dans cette représentation, toute section de balisage, quelle qu'elle soit, *interrompt* le flux textuel. Or, les balises insérées dans le FEW ne se situent pas toutes sur le même pied : à côté de balises structurelles telles que les débuts et fins de paragraphes (<p>, </p>), se trouvent des balises typographiques (indiquant par exemple la mise en grasses ou en italiques d'une portion de texte) qui n'ont aucune portée structurelle et interrompent le flux textuel de manière non prévisible. Les balises entourant les appels de note posent le même problème, car un appel de note peut apparaître n'importe où dans le ruban de caractères.

Les sections de balisage, en interrompant le flux textuel de façon non prévisible, peuvent empêcher la lecture du texte (cf. Tannier 2006) par un algorithme et, notamment, la détection de séquences de caractères spécifiques.

Afin de résoudre ce problème, nous construisons, à partir de la représentation du texte en séquences de *chunks*, une nouvelle représentation qui empêche les balises d'interrompre le flux textuel. Dans cette représentation, les balises peuvent être rendues invisibles, ce qui permet de resolidariser les sections de texte séparées par ces balises et, donc, de conserver, malgré l'insertion de balises XML au fur et à mesure du processus de rétroconversion, des blocs cohérents de discours fewien.

Contrairement à ce qui a déjà été proposé par ailleurs (cf. Tannier 2006), nous ne décidons pas, une fois pour toutes, quelles balises seront conservées et lesquelles seront rendues invisibles dans la totalité du document. Il est, en effet, nécessaire de pouvoir sélectionner finement quelles balises sont rendues invisibles pour chacun des algorithmes, en fonction des besoins. Contrairement à la représentation en séquences de *chunks* qui est permanente du début à la fin de la rétroconversion, cette représentation du texte est donc dynamique, c'est-à-dire qu'elle peut être construite "à la volée" à tout moment par un algorithme. Il s'agit en outre d'une représentation virtuelle, dans le sens où les balises ne sont pas réellement supprimées du document, mais rendues invisibles de façon temporaire, au moment où une opération de lecture du texte le nécessite. La position des balises rendues invisibles est conservée en mémoire.

Nous désignons cette représentation dynamique du texte par le terme de *chaîne (de caractères) virtuelle*. Une chaîne virtuelle est un texte continu, résultant de la concaténation du texte de plusieurs *text chunks* séparés par des *tag chunks* dans le document XML traité. La position de chaque caractère de la chaîne virtuelle dans le document XML sous-jacent est conservée en mémoire et aisément accessible (cf. Briquet et Renders 2010).

Créer une chaîne virtuelle (ou une séquence de chaînes virtuelles) revient à définir, pour une partie du document commençant à un *chunk* spécifié, quels éléments XML (reprenant les balises ouvrantes et les balises fermantes correspondantes) sont laissés visibles et lesquels sont rendus invisibles. La rencontre d'une balise visible met fin à la construction d'une chaîne virtuelle. Afin de rendre ce mécanisme plus intéressant encore, nous ajoutons à ces deux propriétés la possibilité de rendre invisible, non seulement un élément XML, mais son contenu également (que ce contenu se compose de *tag chunks* ou de *text chunks*). Afin de délimiter la partie du document concernée, une (au moins) des balises visibles est spécifiée comme étant terminale, ce qui signifie que des chaînes virtuelles sont construites jusqu'à la rencontre de cette balise dans le texte. En-

fin, nous ajoutons la possibilité de détecter des balises non prévues, ce qui signifie que leur présence dans la partie du document concernée, étant considérée comme anormale, conduira à une alerte. La détection d'une balise inattendue dans une section du texte donnée est un moyen supplémentaire de détecter des erreurs éventuelles (provenant de causes diverses).

Les propriétés définies lors de la création d'une chaîne virtuelle sont donc au nombre de cinq : un élément XML peut être inattendu, terminal, visible (mais non terminal), invisible ou invisible avec son contenu. Nous réserverons le terme d'*invisible* à ce dernier cas et nous parlerons d'élément *transparent* lorsque seules les balises sont invisibles, mais que le contenu de l'élément XML reste visible.

Lors de la construction d'une chaîne virtuelle, toute élément XML se voit attribuer une seule de ces cinq propriétés. Les propriétés sont appliquées au nom de l'élément et concernent donc à la fois la balise ouvrante <x>, la balise fermante </x> et la balise vide <x/>.

Considérons le résultat de l'application de chaque propriété à l'élément x dans l'extrait ci-dessous :

"texte1<x>texte2<y>texte3</y></x><z>texte4</z>texte5"

1. Une balise *terminale* délimite la fin de la chaîne virtuelle. Si l'élément x est défini comme terminal, la chaîne construite à partir de l'extrait se résume à "texte1" si l'on construit la chaîne à partir du début de l'extrait.

2. Une balise *visible* interrompt le flux textuel. Si l'élément x est défini comme visible, les deux sections texte1 et texte2 situées de part et d'autre de la balise ouvrante <x> ne sont pas concaténées. La présence de n balises visibles dans le texte conduit donc à la construction de n ou n+1 chaînes virtuelles.

3. Une balise *transparente* est ignorée, mais le contenu éventuel situé entre la balise ouvrante et la balise fermante correspondante reste visible. Si x est défini comme transparent, les deux sections texte1 et texte2 sont concaténées en une seule section :

   "texte1texte2<y>texte3</y><z>texte4</z>texte5"

   La transparence de l'élément x associée à la visibilité des éléments y et z conduirait donc à la création de quatre chaînes virtuelles : "texte1texte2", "texte3", "texte4" et "texte5".

4. Une balise *invisible* est ignorée avec son contenu : tout ce qui se trouve entre la balise ouvrante et la balise fermante correspondante est ignoré également (qu'il s'agisse de *text chunks* ou de *tag chunks*). L'invisibilité de l'élément x provoquerait donc le résultat (virtuel) "texte1<z>texte4</z>texte5" et conduirait, si l'élément z reste visible, à la création de trois chaînes virtuelles ("texte1", "texte4" et "texte5").

5. Une balise *inattendue* est une balise qui n'est pas censée apparaître dans le texte traité. Si x est défini comme inattendu, sa présence dans l'extrait conduira à une alerte.

4.4.2.4 Choix d'une représentation

Les algorithmes de rétroconversion s'appuient sur la deuxième représentation (séquence de *tag* et *text chunks*) ou la troisième (texte continu résultant de la concaténation de *text chunks*), en fonction du type d'opérations de recherche et de mise à jour souhaitées. Ces deux représentations permettent d'exprimer des raisonnements linguistiques complexes en utilisant un formalisme adapté et facile d'utilisation. Le choix entre les deux représentations du texte s'opérera dans chaque algorithme en fonction des besoins.

4.4.3 Interaction du linguiste avec le système de rétroconversion

Malgré l'attention et le soin portés à la conception des algorithmes de rétroconversion, des variations ou situations imprévues ne manqueront pas de survenir pour un certain nombre d'articles, tant le FEW est rempli de surprises.

Chaque algorithme émet un avertissement dès qu'une situation ambiguë ne peut être résolue automatiquement, par exemple en cas d'appels de note manquants ou en cas de détection d'un élément qui peut relever de deux types d'information différents sans que le contexte ne permette de trancher. Chaque algorithme émet également un avertissement lorsqu'est détecté un passage d'un article méritant une inspection visuelle par un linguiste. Par exemple, un avertissement est émis en cas de non-détection d'un commentaire : l'absence de commentaire dans un article est en effet rare et pourrait donc provenir d'une erreur de l'algorithme. Tous ces avertissements sont consignés à la fin des fichiers XML produits, au sein d'un historique résumant les modifications apportées à l'article par le système de rétroconversion. Tout algorithme détectant une inconsistance grave dans l'article en cours de traitement – y compris parmi les balises déjà insérées – termine son exécution ainsi que celle de la séquence de rétroconversion, et émet un rapport d'échec.

Étant donné que chaque algorithme de rétroconversion émet un fichier XML intermédiaire, le linguiste peut, en cas de problème (soit problème grave détecté automatiquement en temps réel et conduisant à un arrêt non prévu du système, soit problème mineur consigné par le logiciel et examiné a posteriori par le linguiste, soit problème de toute nature détecté a posteriori par le linguiste), déterminer, dans la séquence de fichiers XML générés, le premier fichier XML où l'erreur apparaît, corriger manuellement ce fichier avec un éditeur de texte, puis enfin relancer automatiquement la séquence de rétroconversion à partir de l'algorithme correspondant (cf. Briquet et Renders à paraître). Il va de soi que tous les fichiers XML intermédiaires postérieurs au fichier XML modifié sont remplacés par les fichiers produits par la réexécution de la séquence de rétroconversion.

Pour aider le linguiste à trouver rapidement le fichier XML où apparaît une erreur, un outil Open Source de visualisation de différences entre fichiers textes[11] est automatiquement lancé par le logiciel dès la fin de la rétroconversion. Cet outil permet de visualiser aisément les différences entre les versions successives des fichiers XML produits. Les linguistes peuvent ainsi vérifier le document final ainsi que toutes les étapes intermédiaires de la rétroconversion.

Un processus éditorial distribué a en outre été proposé (cf. Briquet et Renders à

[11] JMeld, `http://sourceforge.net/projects/jmeld/`.

paraître), afin de permettre aux linguistes de collaborer à distance sur le projet de rétroconversion du FEW.

4.5 Conclusion

Le système présenté ici permet de rétroconvertir le FEW article par article. Sans attendre que l'ensemble du FEW soit numérisé ou saisi manuellement, il est possible à un linguiste de sélectionner un ensemble d'articles en fonction d'un projet de recherche particulier, de les préparer de façon à respecter le format d'entrée défini, de les soumettre à validation et de les rétroconvertir, puis éventuellement d'examiner les avertissements émis et de corriger le document. Toutes ces étapes ont été exposées ci-dessus, excepté le contenu des 40 algorithmes qui constituent le noyau de la rétroconversion. Ces algorithmes sont expliqués dans le chapitre suivant.

Chapitre 5

Algorithmes de rétroconversion

5.1 Introduction

Le noyau du logiciel de rétroconversion du FEW est constitué d'une séquence de 37 algorithmes, destinée à être appliquée à chaque article individuellement (→ 4.4.1). Ces algorithmes se subdivisent en trois groupes : algorithmes de prétraitement, de balisage sémantique et de post-traitement.

Les algorithmes *de prétraitement*, au nombre de huit, vérifient que le document soumis au logiciel répond à une série de normes et de propriétés indispensables pour le bon fonctionnement des algorithmes suivants. Leurs tâches vont de la normalisation des caractères d'espacement à la vérification de l'équilibrage des paires de guillemets, en passant par la mise en grasses de l'astérisque précédant un étymon. Ces algorithmes sont les seuls à modifier le contenu textuel du FEW en intervenant parfois directement dans le texte[1].

Les algorithmes *de balisage sémantique*, au nombre de 24, constituent le groupe le plus important. Ils ont pour objectif de reconnaître les divers types d'information du FEW définis précédemment (→ 3) et de les baliser conformément au modèle proposé. Sauf exception, ils ne modifient pas le texte même du FEW : ils se contentent d'y ajouter des balises XML. Par exemple, l'algorithme de balisage *tag-etymon* insère les balises <etymon>...</etymon> autour de chaque étymon.

Les algorithmes *de post-traitement*, au nombre de cinq, clôturent la séquence. Ils analysent le balisage inséré par les algorithmes précédents et en déduisent des informations utiles pour la suite du processus d'informatisation, afin notamment de faciliter l'évaluation du résultat et les corrections manuelles. Les algorithmes de post-traitement remédient donc en partie aux limitations des algorithmes de balisage. Les analyses effectuées par les algorithmes sont conservées en fin d'article, sous forme de commentaires XML.

Le *nom* de chacun des algorithmes a été choisi de façon à expliciter sa finalité.

[1]Ces modifications sont toujours signalées. Un historique des modifications importantes est conservé explicitement. En outre, la totalité des modifications peut être mise en évidence aisément par différenciation d'une version de l'article à l'autre, ce qui permet une inspection complète des modifications apportées automatiquement au FEW (→ 4.4.3).

L'intitulé des algorithmes de balisage commence généralement par *tag-* suivi du type d'information concerné par l'algorithme : *tag-etymon* balise les étymons, *tag-biblio* les sigles bibliographiques etc.

Ce chapitre a pour objectif d'expliquer en détail chacun des 37 algorithmes de rétroconversion. Nous présentons tout d'abord la méthodologie qui a été suivie pour l'élaboration des algorithmes et les principes communs sur la base desquels ils sont construits, ainsi que les conventions d'écriture utilisées. La compréhension des notions expliquées dans cette première section est nécessaire à la lecture des algorithmes. Ces derniers sont ensuite exposés en trois parties, correspondant aux trois groupes distingués ci-dessus. La description formelle des algorithmes se trouve en annexe (→ F).

5.2 Méthodologie de conception d'un algorithme de balisage

Le principe suivi pour l'élaboration des algorithmes de balisage a été de concevoir un algorithme distinct pour chaque type d'information du FEW. Cette façon de procéder n'était pas la seule possible : il eût été envisageable de parcourir le document de façon linéaire, en balisant les types d'information au fur et à mesure qu'on les rencontrait dans le texte. Cette solution serait particulièrement efficace si les éléments du dictionnaire se suivaient dans un ordre attendu et rigoureux, ou si les divers "blocs" d'information pouvaient être repérés avant même de leur attribuer un type, comme ce fut le cas lors de l'informatisation du TLF :

> L'expérience nous a montré qu'il était souhaitable de procéder en deux étapes :
>
> – localiser les frontières des différents objets, sans chercher à identifier tout de suite leur type. Cette étape est essentiellement axée sur la typographie, la ponctuation et la reconnaissance de certains mots-clés,
>
> – la seconde étape, essentiellement axée sur l'analyse du contenu et l'exploitation des règles de succession des éléments, permet d'analyser les différents tronçons définis lors de la première étape et de leur attribuer un type. (Dendien et Pierrel 2003, 16)

Dans le FEW, le grand nombre d'informations implicites (→ 2.6) rend ce procédé non envisageable en pratique. C'est donc type d'information par type d'information que nous balisons un article du FEW, chaque algorithme ajoutant à l'état précédent du texte de nouvelles balises. L'intérêt de cette approche est que nous pouvons, pour la reconnaissance de chaque type d'information, nous servir de tous les autres types d'information préalablement balisés. Dès lors, comme nous le verrons, l'*ordre* dans lequel sont appliqués les différents algorithmes, donc l'ordre dans lequel sont balisés les différents types d'information, joue un rôle considérable dans le succès de la rétroconversion.

La méthodologie suivie pour élaborer un algorithme de balisage comprend cinq étapes, détaillées ci-dessous. La première étape consiste à examiner la version imprimée du FEW afin d'établir des indicateurs de reconnaissance, c'est-à-dire des critères

permettant de reconnaître le type d'information qui doit être balisé par l'algorithme. L'étape suivante consiste à traduire ces indicateurs en termes informatiques, c'est-à-dire à transformer l'énoncé de propriétés conformes à la version imprimée du FEW en leur équivalent dans la version informatique traitée par l'algorithme. La troisième étape consiste à écrire l'algorithme proprement dit qui exploitera ces indicateurs. À ce stade, l'algorithme peut être implémenté dans un langage informatique et testé, afin de vérifier qu'il fonctionne de la façon attendue. La validation de l'algorithme s'effectue par son application à un corpus représentatif d'articles du FEW au format FFML (→ 4.3). Cette quatrième étape est idéalement effectuée non seulement pour chaque algorithme séparément, mais aussi avec l'ensemble des algorithmes précédents dans le séquençage envisagé, afin de se rapprocher le plus possible du traitement final. L'analyse des résultats individuels et collectifs permet, dans une cinquième et dernière étape, de détecter les problèmes éventuels et de tenter d'y remédier par divers moyens, en modifiant par exemple le contenu de l'algorithme, ou sa place dans la séquence.

5.2.1 Première étape : identification des indicateurs de reconnaissance

5.2.1.1 Définition

Dans le chapitre 2.6.3, nous avons mis en évidence les différences entre un humain et une machine dans la résolution de l'implicite fewien. Dans ce contexte, nous entendions par *implicite* un contenu informationnel absent en structure de surface et que le lecteur humain pouvait rétablir au moyen de ses connaissances grammaticales, lexicales ou encyclopédiques. Les informations présentes en structure de surface, telles que l'entrée d'un article, l'étymon-vedette ou un numéro de note, n'ont pas été abordées dans ce chapitre. Pourtant, il n'est dit nulle part dans le FEW que ce qui est imprimé en grasses dans une taille de police plus grande est la vedette de l'article : ce fait relève des connaissances encyclopédiques du lecteur. Il est donc implicite et est dès lors, lui aussi, susceptible de poser problème dans le cadre d'une analyse informatisée.

Lorsqu'il est question d'une machine et non plus d'un lecteur humain, la définition de l'implicite est donc à élargir sensiblement : elle dépasse le « contenu informationnel » pour englober le « type d'information ». Ce dernier ne pose généralement pas de problème de reconnaissance à un humain habitué à la consultation de dictionnaires. Le lecteur du FEW, même non initié, reconnaît instantanément une note, une définition, une vedette ou une citation. En revanche, il va de soi qu'une machine ne peut déterminer intuitivement si un extrait de texte est une vedette ou une citation. Si l'on veut qu'elle puisse reconnaître tel type d'information, il faut lui donner des critères de reconnaissance, c'est-à-dire des propriétés observables, par exemple la mise en grasses, la présence d'un caractère de ponctuation spécifique etc. Nous appellerons *indicateur de reconnaissance* toute propriété qui permet de reconnaître un type d'information déterminé.

5.2.1.2 Typologie des indicateurs

Il s'agit donc de chercher ce qui fait la *particularité* de chaque type d'information que nous voulons baliser. Ces particularités (ou propriétés) peuvent être *internes* au discours fewien (nous parlerons de *propriétés linguistiques*) ou *externes*, c'est-à-dire

relevant de l'inscription sur papier du discours lexicographique (traits relevant notamment de la mise en forme et de la mise en page ; nous parlerons de *propriétés typographiques*).

Puisque c'est en fin de compte une machine qui recherchera mécaniquement les indicateurs, ces derniers doivent être observables. Parmi les propriétés linguistiques, les propriétés *sémantiques* (par exemple le fait qu'un étymon puisse être défini comme la base d'une famille lexicale) sont sans utilité. En revanche, les propriétés relevées peuvent être des propriétés *morphologiques* (l'astérisque dans le FEW peut être considéré comme un morphème typiquement accolé à un étymon), *syntaxiques* (l'étymon occupe une position particulière dans la structure du FEW) ou *lexicales* (l'étymon relève d'une catégorie particulière de mots dont on peut constituer, du moins dans le cadre du FEW, un lexique).

Ces exemples montrent que le FEW présente un discours analysable comme tout discours en langue, la langue du FEW pouvant être appréhendée comme une *langue spéciale*. Notre analyse du FEW pourra donc utiliser, quand ce sera nécessaire pour les besoins de l'exposé, l'appareil terminologique utilisé en linguistique pour décrire toute langue et toute variété de langue.

Si nous affinons l'analyse des propriétés linguistiques, nous constatons, sans étonnement, que les *propriétés morpho-lexicales* permettent souvent de déterminer la nature d'un élément du discours fewien, tandis que les *propriétés syntaxiques* ont à voir avec sa fonction. Un même étymon, par exemple, peut occuper dans le FEW diverses fonctions : vedette de l'article, sous-lemme, étymon de renvoi (fonctions qui pourront constituer autant de sous-types pour le relevé des indicateurs). C'est essentiellement la place de l'étymon dans la syntaxe du FEW qui permet de distinguer entre ces fonctions, alors que son appartenance au lexique des étymons et la présence éventuelle d'un astérisque permettent de le classer dans la catégorie des étymons. Dans les chapitres suivants, nous utiliserons les termes de *nature* et de *fonction* avec ces acceptions lorsque nous les appliquerons à un type d'information du FEW.

Les indicateurs seront qualifiés de *textuels* lorsqu'ils relèvent de la morphologie ou du lexique et concernent la nature d'une information, de *positionnels* lorsqu'ils relèvent de la syntaxe. Les indicateurs typographiques constituent un groupe particulier qui met en évidence, selon les cas, des propriétés soit syntaxiques, soit morpholexicales. Notons que dans le cadre de la rétroconversion du TLF, Dendien et Pierrel 2003 avaient proposé une typologie semblable des indicateurs en distinguant la typographie, le contenu textuel et la succession des éléments (contexte dans lequel ils apparaissent).

5.2.1.3 Spécificité et fiabilité

Un indicateur de reconnaissance doit idéalement répondre à deux exigences : sa spécificité et sa fiabilité. La spécificité est le fait que l'indicateur ne s'applique pas à d'autres types d'information. Sans cette exigence de spécificité apparaîtraient des *faux positifs*, c'est-à-dire le balisage d'éléments qui vérifient l'indicateur alors qu'ils ne relèvent pas du type d'information concerné. La fiabilité est le fait que toutes les occurrences du type d'information concerné contiennent cet indicateur. Sans cette exigence de fiabilité apparaîtraient des *faux négatifs*, c'est-à-dire des éléments non balisés parce qu'ils ne vérifient pas cet indicateur, alors qu'ils relèvent bien du type d'information concerné.

Ces deux exigences théoriques se heurtent à la réalité complexe du « terrain », à savoir le texte fewien. En pratique, peu de types d'information bénéficient d'indicateurs qui leur soient totalement spécifiques. Le plus souvent, un indicateur (par exemple le rendu typographique en italiques) est partagé par plusieurs types d'information différents (par exemple les lexèmes, les citations, ainsi que les étymons dans certains contextes). De même, les nombreuses incohérences du FEW empêchent de considérer tout indicateur comme fiable sur la totalité du dictionnaire. Heureusement, la non-spécificité et la non-fiabilité de certains indicateurs ne signifient pas la non-faisabilité de la rétroconversion. En effet, si un indicateur pris séparément est rarement suffisant à l'identification d'un type d'information, une combinaison bien choisie et bien articulée d'indicateurs lève presque toujours le doute.

Un des objectifs de cette thèse est précisément de *définir pour chaque type d'information une combinaison d'indicateurs qui puisse répondre aux exigences de spécificité et de fiabilité requises*. Une information sera reconnaissable, donc balisable de façon automatique, uniquement si elle peut être définie par une combinaison de propriétés qui ne s'applique qu'à elle. Certains types d'information regroupant en réalité plusieurs sous-types, c'est parfois plusieurs combinaisons d'indicateurs qui devront être définies : par exemple, la reconnaissance des étymons (effectuée par un même algorithme) met en jeu des indicateurs différents selon qu'il s'agit du sous-type des étymons-vedettes ou des étymons cités dans le corps des articles.

5.2.2 Deuxième étape : traduction en termes informatiques

Les algorithmes de rétroconversion ne s'appliquent pas directement au texte imprimé du FEW. Une fois définies, les propriétés spécifiques à chaque type d'information doivent être traduites en langage informatique afin de correspondre à la version électronique du texte.

La représentation du texte fewien traitée par un algorithme a été décrite précédemment (→ 4.4.2) : il s'agit d'une chaîne de caractères codés en Unicode, dans laquelle se trouvent des balises XML (dont toutes les balises insérées par les algorithmes déjà appliqués précédemment). Traduire une propriété en termes informatiques revient donc à y assimiler des codes Unicode ou des balises XML. Par exemple, la mise en grasses (propriété typographique) est représentée sous la forme de balises ... ; la propriété syntaxique « précédé d'un tiret » doit être traduite par « précédé d'un des trois caractères Unicode "-" (U+002D), "–" (U+2013) ou "—" (U+2014) ».

Puisque c'est très souvent une combinaison d'indicateurs (et non un seul) qui permet de reconnaître avec fiabilité et spécificité un type d'information donné, l'opération de traduction se présente souvent sous la forme d'une séquence de caractères et de balises. Par exemple, un étymon non attesté pourrait être reconnu par la présence d'un astérisque (caractère Unicode U+002A) situé juste avant une balise , l'astérisque et la balise devant se trouver entre une balise <entry> et une balise </entry>, ce qui donne la séquence suivante : <entry>. . .*etymon. . .</entry>.

Reconnaître un type d'information revient dès lors à détecter la présence de séquences spécifiques de caractères (lettres, chiffres, signes de ponctuation) et de balises. Ces séquences qui ont pour nous un *sens* et dont nous cherchons à vérifier la *présence* dans le texte sont appelées ici *motifs* (traduction courante de l'anglais *pattern* utilisé en informatique pour désigner la même notion). Par exemple, le mot *completus*, un asté-

risque devant une balise <b>, ou encore une suite de quatre chiffres formant une date située entre 1000 et 2000, constituent autant de motifs qui peuvent être recherchés dans un texte.

Dans le cadre de la rétroconversion du FEW, la détection des motifs se fait au moyen de deux outils différents, selon que le motif est connu de façon exacte ou non : ces outils sont les listes de mots-clés et les expressions régulières. L'utilisation de ces outils pose en pratique quelques problèmes, dus à la complexité du FEW autant qu'à l'utilisation du format XML. Nous exposons ci-dessous l'utilisation qui est faite de ces outils, les problèmes rencontrés et les méthodes de résolution qui ont été choisies.

5.2.2.1 Listes de mots-clés

Une liste de mots-clés est un ensemble de mots ou de suites de caractères qui partagent une particularité identifiée comme rentable pour le balisage d'un type d'information du FEW. Par exemple, il est possible de constituer un lexique des étymons du FEW, ou encore une liste des noms des rédacteurs. Nous parlerons dans la suite de *mot-clé*, étant entendu qu'il peut s'agir d'un ensemble de mots, d'une locution ou même d'une phrase.

Pour la rétroconversion du FEW, 26 listes ont été établies, chacune concernant un type d'information particulier. Leur nom commence par *few-* et se termine par *-base*. *Few-etymon-base* par exemple contient la liste des mots-clés correspondant aux étymons, *few-geoling-base* la liste des étiquettes géolinguistiques.

Nous nous dotons d'une fonction de recherche exacte de mots-clés d'une liste donnée opérant sur la représentation informatique du texte fewien. La particularité de cette recherche est qu'elle ne permet aucune *variante* : seule une correspondance exacte, caractère par caractère, entre un mot-clé d'une liste donnée et une partie de texte du FEW mènera à un résultat positif. Cela implique que le codage des caractères, y compris spéciaux, doit être rigoureusement identique dans le texte du FEW et dans la liste. Il est important également que les incohérences du FEW soient prévues : ainsi, un algorithme basé uniquement sur la fonction de recherche exacte ne pourra reconnaître *Smiřicky* dans le texte si la liste comporte uniquement *Smiřický*. De même, il ne pourra reconnaître *Cham-bon* (avec un tiret de coupure de ligne) à partir du mot-clé *Chambon*. Les cas de variation arbitraire de ce genre doivent être soit intégrés dans les listes, soit résolus par d'autres moyens algorithmiques.

5.2.2.2 Expressions régulières

Une expression régulière est un moyen de représenter une séquence de caractères en y laissant des possibilités de variantes (cf. Kleene 1956) ; cette séquence décrit donc un *ensemble de formes possibles*. Les expressions régulières peuvent représenter des séquences de caractères dont on ne peut, ou ne veut, établir la liste. Par exemple, plutôt que de lister toutes les dates comprises entre 1000 et 2000, nous pouvons les représenter de façon plus concise par l'expression régulière suivante : [1][0-9]3|2000 qui signifie « soit le chiffre 1 suivi de trois chiffres entre 0 et 9, soit le nombre 2000 ».

Les expressions régulières constituent un formalisme puissant [2]. La lecture et la

[2]Cf. Stubblebine 2003 ; pour une description formelle complète de la syntaxe des expressions régulières,

compréhension des algorithmes décrits aux pages suivantes nécessite de connaître quelques-uns des symboles utilisés dans les expressions régulières. Signalons notamment les opérateurs suivants :

– la barre verticale | , signifiant « ou » ;

– le signe ?, indiquant que l'élément qui précède apparaît 0 ou 1 fois ;

– l'astérisque *, indiquant que l'élément qui précède apparaît 0, 1 ou plusieurs fois ;

– le signe +, indiquant que l'élément qui précède apparaît 1 ou plusieurs fois ;

– les accolades {x} ou {x-y}, indiquant que l'élément qui précède apparaît x fois ou entre x et y fois :{3}, {3-5}.

Par élément, nous désignons un caractère ou un ensemble de caractères. Un ensemble de caractères se trouve entre parenthèses. Les crochets sont quant à eux utilisés pour spécifier plusieurs caractères possibles : par exemple, l'expression [ab] signifie « a ou b » ; [aepq] « a ou e ou p ou q » ; l'expression [a-d] équivaut à [abcd].

5.2.2.3 Difficultés de détection des motifs fewiens et méthodes de résolution

Gestion des caractères spéciaux

Les caractères spéciaux du FEW posent un problème dans l'écriture d'expressions régulières. Par exemple, une expression régulière possible pour définir un « mot » est généralement `/\s[A-Za-z]+\s/`, signifiant « une ou plusieurs lettres appartenant à l'ensemble des lettres de a à z en minuscules ou en majuscules, délimitées par des espaces ». Dans le FEW, cette définition n'est pas applicable : en effet, il faut ajouter à cet ensemble tous les caractères spéciaux (notamment phonétiques) qui peuvent apparaître au sein d'un mot. Le nombre de caractères spéciaux présents dans le FEW rend non envisageable en pratique l'écriture d'une expression régulière concise et utilisable.

Une solution à ce problème consiste à définir pour chaque caractère spécial du FEW une version « aplatie », c'est-à-dire un caractère standard utilisable dans les expressions régulières qui en représente en quelque sorte une version neutralisée, non diacritée. Par exemple, le caractère phonétique « ǫ̆ » est aplati en « o ». Les versions aplaties des caractères spéciaux sont consignées dans la table de caractères du FEW (→ 4.2.2) utilisée par les algorithmes. Il est dès lors possible, lors de la détection d'un motif, d'aplatir le texte du FEW avant d'y appliquer une expression régulière. Cette méthode permet d'écrire les expressions régulières de façon standard sans y intégrer les nombreux caractères spéciaux.

Gestion des balises parasites

Les listes de mots-clés et les expressions régulières doivent correspondre au texte du FEW et dépendent donc de la représentation informatique du texte fewien. Parmi les

---

voir The Open Group, 1997. Regular Expressions. The Single UNIX Specification, Version 2. `http://www.opengroup.org/onlinepubs/007908799/xbd/re.html`.

trois possibilités de représentation de ce texte que nous avons définies précédemment (→ 4.4.2), il s'agit de choisir celle qui offre le plus de confort et d'efficacité.

Prenons un exemple construit pour les besoins de l'exposé : la recherche d'un sigle bibliographique connu (SIGLE), qui peut être suivi d'une date et d'un numéro d'édition en exposant. Dans la version informatique du texte fewien, il s'agit de rechercher un mot-clé (sigle bibliographique) suivi éventuellement d'une séquence de quatre chiffres (date), d'une balise <e>, d'un nombre et d'une balise </e> (numéro d'édition en exposant). Une expression régulière « naïve » permettant de détecter un tel sigle ainsi que la date et le numéro d'édition qui l'accompagnent serait par exemple « SIGLE ([0-9]4<e>[1-9]+</e>) ? ».

Cette expression régulière est naïve, car elle se fonde sur le texte du FEW tel qu'il est fourni dans un fichier FFML (→ 4.3). Elle ne tient pas compte des balises éventuellement insérées entre-temps dans le texte fewien par les algorithmes de rétroconversion. Par exemple, il se peut qu'un algorithme précédent ait déjà inséré les balises <date>…</date> autour de la date. De même, nous savons qu'à tout endroit du texte fewien peuvent apparaître des appels de note. Si ces derniers ont déjà été balisés par un algorithme précédent, un élément XML <appelnote>...</appelnote> peut apparaître n'importe où dans la séquence de caractères recherchée. Ces balises parasites risquent d'empêcher la détection du sigle par l'expression régulière.

L'expression régulière doit donc être écrite de façon à tenir compte des balises parasites. Or, les représentations informatiques du texte fewien définies précédemment (→ 4.4.2) permettent différentes façons de traiter ces balises parasites : l'expression régulière sera donc différente selon la représentation choisie.

Dans une représentation du FEW comme simple chaîne de caractères, les balises ne sont pas distinguées du texte fewien : elles sont dès lors à intégrer en tant que caractères textuels dans l'expression régulière. Par exemple, un appel de note préalablement balisé peut être « traduit » en "<appelnote><e>[1-9][0-9]*</e>)</appelnote>". L'appel de note pouvant apparaître n'importe où, il est à intégrer (entre parenthèses marquées d'un point d'interrogation, signifiant que l'ensemble est facultatif) à chaque endroit de l'expression régulière susceptible d'être concerné : « SIGLE(<appelnote><e>[1-9]+</e>) </appelnote>) ? [0-9]{4}<e>[0-9]*</e> (<appelnote><e>[1-9]+</e>) </appelnote>) ? »

La nécessité de prévoir tout endroit où peut apparaître un élément parasite (appel de note ou autre) rendrait l'écriture des expressions régulières très complexe. La représentation du FEW comme simple chaîne de caractères n'est donc pas optimale à cet égard.

La deuxième représentation, qui distingue les balises (*tag chunk*) du reste du texte (*text chunk*), apporte une amélioration en permettant d'extraire de l'expression régulière toutes les balises. Elle présente toutefois l'inconvénient majeur de découper le texte : une expression régulière ne peut plus s'appliquer qu'à une section de texte située entre deux balises. Les balises parasites posent dès lors un problème en créant des interruptions non signifiantes. Une solution serait de regrouper les portions de texte ainsi découpées en une seule longue chaîne. Un problème subsiste cependant : comment conserver les balises signifiantes, par exemple la balise <e> signalant un exposant ?

La représentation du texte sous forme de chaîne virtuelle (→ 4.4.2.3) apporte une solution qui évite ce problème. La possibilité de donner aux balises des propriétés différentes de visibilité permet, dans l'exemple ci-dessus, de créer une chaîne virtuelle dans

laquelle les balises <appelnote> sont *rendues invisibles avec leur contenu*. L'écriture des expressions régulières est dès lors simplifiée, puisqu'elle ne doit plus tenir compte des appels de note : si un appel de note apparaît à n'importe quel endroit du texte, il sera tout simplement ignoré comme s'il était absent. La représentation virtuelle de la séquence

SIGLE<appelnote> ... </appelnote> 1954<e>2</e>

devient en effet :

SIGLE 1954<e>2</e>

Nous pourrions également décider de rendre les balises <e>...</e> *transparentes* dans la chaîne virtuelle (le numéro d'édition restant alors visible). Dans le cas de sigles dont la date et le numéro d'édition sont connus, ce procédé permet d'intégrer ce sigle suivi de sa date et son numéro d'édition tel quel dans une liste de mots-clés (par exemple SIGLE 19542).

Cette possibilité d'effacer temporairement les balises problématiques simplifie énormément la construction des algorithmes en offrant des solutions élégantes à des problèmes parfois complexes. Le principe suivi est de *complémenter la puissance des expressions régulières par la flexibilité des chaînes virtuelles*, de manière à pouvoir raisonner de manière naturelle sans avoir recours à une syntaxe très lourde, ingérable en pratique.

Concrètement, chaque fois que nous chercherons un motif dans le texte fewien (que ce soit à l'aide d'une liste de mots-clés ou d'une expression régulière), nous construirons préalablement une chaîne virtuelle[3], dans laquelle tous les éléments XML auront reçu une des cinq propriétés définies plus haut (→ 4.4.2.3). L'attribution aux éléments XML de l'une des cinq propriétés est spécifiée dans les algorithmes sous l'intitulé *partition*, de la façon suivante :

```
partition()
balises terminales: p
balises transparentes: b, i, sc, e, lang_etymon
balises invisibles: appelnote
balises visibles: par défaut
balises inattendues: aucune
```

Par convention, dans les partitions, "x" désigne à la fois la balise ouvrante <x>, la balise fermante </x> et la balise vide <x/>.

Gestion des collisions et enchâssements

Des problèmes peuvent se produire lorsque plusieurs listes comportent des mots-clés identiques. Par exemple, *Chambon* est dans le FEW soit le nom d'un rédacteur (mot-clé consigné dans *few-signature-base*), soit le nom d'une localité située en Lozère (mot-clé consigné dans *few-geoling-base*). Ce mot-clé risque d'être balisé deux fois différemment, par deux algorithmes utilisant chacun une des deux listes, de la façon suivante :

<geoling><signature>Chambon</signature></geoling>.

La plupart de ces collisions entre des mots-clés identiques appartenant à des listes

[3]L'expression "construire une chaîne virtuelle" pourra désigner, en pratique, la création d'une séquence de plusieurs chaînes virtuelles (→ 4.4.2.3).

différentes peuvent être résolues par un séquençage judicieux des algorithmes, combiné à la possibilité de rendre invisibles certains éléments XML dans la création des chaînes virtuelles. Le séquençage des algorithmes permet en effet qu'un item ambigu, appartenant à la fois à une liste X et à une liste Y utilisées par deux algorithmes différents A et B, soit toujours identifié selon le premier des deux algorithmes qui apparaît dans la séquence (l'algorithme A). Le mécanisme des chaînes virtuelles (→ 4.4.2.3) permet ensuite, dans l'algorithme B, d'effacer temporairement les éléments XML problématiques, afin qu'ils ne soient pas balisés. Selon ce procédé, les items ambigus sont uniquement balisés par l'algorithme A et ne sont pas reconnus par l'algorithme B.

Cette solution est élégante lorsque l'utilisation de la liste X par l'algorithme A est combinée à d'autres critères de reconnaissance qui permettent de s'assurer que les items ambigus repérés par l'algorithme A relèvent bien de l'information concernée par ce dernier. Lorque les critères de reconnaissance utilisés par les algorithmes ne permettent pas de lever les ambiguïtés, la seule solution consiste à notifier le caractère ambigu de ces items, pour inspection et vérification par un expert. Dans ce cas, l'algorithme A, tout en les balisant, les marque explicitement comme ambigus et émet un avertissement demandant une vérification manuelle. C'est la solution qui a par exemple été choisie pour les collisions entre les références bibliographiques et les étiquettes géolinguistiques (→ 5.4.9, 5.4.10).

D'autres problèmes peuvent se produire en cas d'enchâssements, c'est-à-dire lorsqu'un mot-clé d'une liste est inclus dans un mot-clé d'une autre liste[4]. Ces situations peuvent amener les mêmes problèmes de double balisage que les situations de collision, à la différence que l'un des deux balisages contient une portion de texte supplémentaire. Par exemple, l'enchâssement de « d » (mot-clé de la liste des marqueurs alphanumériques, balisé <pnum>) dans « d. » (mot-clé de la liste des langues [= *deutsch* « allemand »], balisé <lang>) peut conduire au balisage suivant :

```
<lang><pnum>d</pnum>.</lang>
```

Les cas d'enchâssements sont résolus de la même façon que les cas de collisions.

5.2.3 Troisième étape : écriture des algorithmes

Un algorithme sera ici commodément défini comme une séquence d'instructions qui, suivie par une machine, mène au résultat demandé. Dans la rétroconversion du FEW, le résultat attendu est, pour la plupart des algorithmes, le balisage d'un type d'information particulier. La séquence d'instructions décrit précisément comment utiliser les indicateurs de reconnaissance pour atteindre cet objectif. Dans les chapitres suivants, les *objectifs* précis de chaque algorithme seront toujours spécifiés avant l'explication de son fonctionnement, afin de souligner la préséance de l'utile sur le faisable qui a guidé notre démarche.

La séquence d'instructions peut être représentée dans un langage de programmation ou dans un pseudo-langage défini par l'algorithmicien. Dans ce dernier cas, on suppose l'existence d'une machine abstraite. Dans cette thèse, les algorithmes sont présentés dans un tel pseudo-langage, inspiré librement du *métalangage des commandes gardées* de Dijkstra (Dijkstra 1976).

[4]Les enchâssements entre mots-clés d'une même liste ne sont pas problématiques, car lors de la détection des mots-clés d'une même liste, le mot-clé le plus long a toujours priorité sur le mot-clé le plus court.

L'utilisation d'un métalangage présente un avantage par rapport à un langage de programmation : il permet de décrire l'essentiel du raisonnement sans nécessairement entrer dans des détails d'implémentation. Dans des situations d'interdisciplinarité où linguistes et informaticiens coopèrent, le métalangage s'avère un excellent moyen de communication. Il permet au linguiste de décrire de façon rigoureuse un processus de résolution tout en laissant à l'informaticien, qui doit implémenter ce processus, une liberté plus ou moins grande concernant les détails techniques (choix du langage de programmation, choix de certaines structures de données etc.). Chacune des deux disciplines reste ainsi souveraine dans son domaine d'expertise propre, même si leur interaction peut conduire à des compromis qui influencent les décisions prises par l'une et l'autre (les choix architecturaux décidés par l'informaticien peuvent, par exemple, amener le linguiste à modifier certains algorithmes pour qu'ils intègrent de nouvelles structures de données).

La description formelle, dans le métalangage choisi, des algorithmes de rétroconversion se trouve en annexe (→ F). Dans les pages suivantes, nous expliquerons le fonctionnement des algorithmes de façon non formelle. Nous décrirons certaines opérations de façon formelle uniquement lorsque cela nous semblera utile pour la compréhension de l'exposé. Le lecteur trouvera en annexes (→ F.1) quelques explications sur le métalangage utilisé.

5.2.4 Quatrième et cinquième étapes : validation et optimisation

Il est difficile de déterminer si un algorithme fonctionne comme prévu (et donc atteint son but) avant de l'avoir implémenté et testé, non par manque de rigueur dans le raisonnement algorithmique, mais parce que le problème à résoudre n'est lui-même pas défini avec certitude (puisqu'un modèle complet du FEW réel n'est pas disponible). Par ailleurs, il est risqué en pratique de vouloir écrire des algorithmes qui se basent sur d'autres algorithmes lorsque ces derniers n'ont pas été préalablement validés ou au moins testés.

Une implémentation simultanée des algorithmes n'a toutefois pas été possible dans les premiers temps de leur élaboration. L'écriture des algorithmes s'est donc d'abord faite de façon théorique à partir de l'analyse du FEW, dans une optique de sécurité : l'objectif était, en l'absence de tests permettant une évaluation des résultats au fur et mesure de l'avancement de la recherche, de spéculer le moins possible. Ceci a conduit à une certaine prudence et, notamment, à la décision d'utiliser de façon prioritaire des moyens de reconnaissance fiables tels que les listes de mots-clés, facilement améliorables, plutôt que de chercher à tout prix à rendre les algorithmes capables de plus de discernement.

L'implémentation des algorithmes et la validation sur corpus de test ont ensuite été réalisées grâce à l'engagement par l'ATILF d'un chercheur en informatique, le Dr Cyril Briquet. Le programme Java complet se trouve en annexe (→ G)[5]. Les résultats de ces tests ont pu conduire pour certains algorithmes à des optimisations, qui sont signalées le cas échéant dans les pages suivantes. Outre la modification de l'algorithme même, trois autres moyens ont été mis en œuvre pour résoudre les problèmes rencontrés :

[5]Le code Java n'utilise pas nécessairement les mêmes noms de variables que le métalangage exposé dans cette thèse ; toutefois, les noms des algorithmes, des listes de mots-clés, des balises et de leurs attributs sont strictement identiques.

1. l'ajout de données dans les listes de mots-clés utilisées par l'algorithme ;
2. l'écriture d'un nouvel algorithme de prétraitement ;
3. la modification de la place de l'algorithme dans la séquence de rétroconversion.

Les pages suivantes présentent les algorithmes dans leur version définitive, avec les optimisations éventuellement apportées après validation sur corpus.

5.3 Algorithmes de prétraitement

Les huit algorithmes de prétraitement sont appliqués au tout début de la séquence de rétroconversion, après la validation XML de l'article au format FFML (→ 4.3.3) et immédiatement après le chargement de l'article dans la représentation en séquences de *chunks* (→ 4.4.2.2). Dans cette représentation du texte, les algorithmes vérifient certaines propriétés, corrigent des erreurs ou signalent des incohérences non détectables par la validation XML. Ils vont plus loin que cette dernière, dans le sens où ils corrigent parfois le FEW lui-même. La raison d'être de ces huit algorithmes est qu'ils conditionnent le succès des algorithmes de balisage.

L'acquisition des articles du FEW au format FFML pouvant a priori se faire autant par saisie manuelle que par numérisation et océrisation, les algorithmes de prétraitement ont été pensés de façon à remédier à des problèmes qui pouvaient provenir de chacun de ces deux modes d'acquisition du texte. Plusieurs de ces problèmes, présentés dans les pages suivantes, ont été réellement détectés sur le corpus de test, composé d'articles pour la plupart d'abord numérisés et océrisés, puis corrigés manuellement.

Parmi les huit algorithmes exposés ci-dessous dans trois sections, les cinq premiers (*detect-corrupted-entities*, *streamline-p-extreme-spaces*, *streamline-void-tags*, *detect-dubious-spacing* et *streamline-breaks*) détectent et corrigent des erreurs de saisie ou de numérisation. Les deux suivants, *streamline-layout-tags* et *streamline-quotes*, corrigent des incohérences qui peuvent provenir du FEW lui-même. Une troisième section contient un seul algorithme au caractère un peu particulier. L'algorithme *merge-split-words* se rapproche en effet des algorithmes de balisage sémantique, puisqu'il analyse le texte du FEW en tentant de déterminer – dans une certaine mesure (et non de manière absolue) – le sens des tirets de fin de ligne.

Parmi ces algorithmes, seul le fonctionnement de *streamline-quotes* et de *merge-split-words* fait l'objet d'une description formelle en annexe (→ F.5, F.2).

5.3.1 Algorithmes de détection d'erreurs de saisie ou de numérisation

5.3.1.1 Algorithme *detect-corrupted-entities*

L'algorithme *detect-corrupted-entities* a pour objectif de détecter, sans les corriger, des erreurs de syntaxe dans la saisie des entités XML définies pour la transcription des caractères spéciaux du FEW. Ces entités commencent toujours par le caractère "&" suivi soit du caractère "#" et d'un code Unicode (par exemple dans "&#x231C ;", entité

désignant le semi-crochet ouvrant ⌜ dont le code Unicode est U+231C) soit, pour les caractères non Unicode, de "few-" suivi d'un nom composé (par exemple "&few-o-ouvert-long ;").

L'algorithme ne vérifie pas que l'entité appartient bien à l'ensemble des entités XML définies pour le FEW, puisque cette vérification est effectuée implicitement par la résolution des entités lors de la validation XML, ni qu'elle correspond bien au caractère du FEW, vérification que seul un humain pourrait assurer par inspection visuelle. En revanche, il détecte les erreurs du type "&x231C ;" ou "#x231C". En effet, ces entités incomplètes (il leur manque soit le # soit le &) ne sont pas reconnues comme entités lors de la validation XML et ne sont donc pas résolues. De même, l'algorithme détecte les entités du type "&few-... ;" qui n'auraient pas été reconnues lors de la validation XML.

Ces erreurs de saisie ne nuisant pas au bon fonctionnement des algorithmes suivants, elles ne sont pas corrigées automatiquement ; seul un avertissement, purement informatif, est émis pour signaler le problème.

5.3.1.2 Algorithme *streamline-p-extreme-spaces*

L'algorithme *streamline-p-extreme-spaces* est appliqué au contenu des seuls éléments XML <few>, <p> et <notes>. À l'intérieur de l'ensemble de caractères situés entre la balise ouvrante et la balise fermante correspondante, il supprime les espaces situés en tête de la chaîne de caractères (donc directement à droite de la balise ouvrante), ainsi que les espaces situés en fin de chaîne (donc directement à gauche de la balise fermante).

Il est à noter que lors du chargement d'un article dans sa représentation en *tag chunks* et *text chunks*, les sauts de lignes, tabulations et autres caractères d'espacement sont remplacés par des espaces, et les espaces redondants sont compactés, de sorte que toute séquence de caractères d'espacement se réduit à un seul espace. En fin de rétroconversion, des sauts de ligne sont ajoutés automatiquement, suivant des règles de formatage définies, lors de l'exportation de la représentation intermédiaire en un fichier XML.

Outre le contrôle de la mise en page, ces opérations garantissent que la version informatique du texte fewien ne contient aucun caractère d'espacement excédentaire, ce qui est nécessaire pour certains algorithmes de balisage qui effectuent des vérifications sur les caractères en début de paragraphe, comme *tag-numbering* (→ 5.4.12) ou *tag-notes* (→ 5.4.1).

5.3.1.3 Algorithme *streamline-void-tags*

L'algorithme *streamline-void-tags* détecte et corrige les éléments XML sans contenu (<x></x>) ou contenant uniquement un caractère d'espacement (<x> </x>). De telles erreurs sont typiques d'un balisage automatisé (par exemple via des macros appliquées en fin du processus d'océrisation) des italiques et des grasses, où les espaces sont balisées comme n'importe quel autre caractère.

La non-correction de ces erreurs poserait un problème pour l'écriture des algorithmes suivants, car ces derniers devraient alors vérifier systématiquement le contenu

des éléments XML avant toute opération concernant ceux-ci. Les éléments XML ne possédant aucun contenu sont directement supprimés par *streamline-void-tags*. S'ils contiennent uniquement un caractère d'espacement, celui-ci est conservé (il fait en effet partie intégrante du texte du FEW), mais les balises qui l'entourent sont supprimées.

5.3.1.4 Algorithme *detect-dubious-spacing*

L'algorithme *detect-dubious-spacing* vérifie la disposition des espaces autour des balises. Les balises FFML sont en effet censées être directement jointes au texte qu'elles entourent, c'est-à-dire que la balise ouvrante <X> est accolée au texte qui la suit, tandis que la balise fermante </X> est accolée au texte qui la précède, comme suit : "texte <X>texte</X> texte". Seule la balise <e>, signalant un exposant, échappe à cette vérification, puisqu'un exposant peut légitimement être collé au texte qui le précède sans aucun espace intermédiaire ("texte<e>1</e> texte").

L'algorithme émet un avertissement s'il détecte un espace juste après une balise ouvrante (<x> texte) ou juste avant une balise fermante (texte </x>), car il est très probable que la balise a été placée par erreur du mauvais côté du caractère d'espacement. L'absence d'espace ou de ponctuation avant une balise ouvrante (texte<x>...) ou après une balise fermante (...</x>texte) fait également l'objet d'un avertissement.

Les avertissements émis par *detect-dubious-spacing* sont purement informatifs, car les manquements ne doivent pas stopper le processus de rétroconversion. En effet, la présence ou l'absence erronée d'espace ou de ponctuation autour des balises n'a généralement pas d'incidence néfaste sur le fonctionnement des algorithmes de balisage. Seules les erreurs autour des balises <few>, <p> et <notes> doivent absolument être corrigées : c'est la raison d'être de l'algorithme *streamline-p-extreme-spaces* (→ 5.3.1.2), qui est appliqué avant *detect-dubious-spacing* (→ 5.3.1.4) dans la séquence de rétroconversion.

5.3.1.5 Algorithme *streamline-breaks*

L'algorithme *streamline-breaks* effectue cinq vérifications à propos des balises <lb/> (laquelle marque le passage à la ligne) et <col/> (laquelle marque le passage à la colonne suivante). Les trois premières visent des normalisations qui consistent en la suppression de balises ou de caractères d'espacement :

1. Suppression des balises <lb/> qui se trouvent entre des paragraphes. En effet, une balise </p> implique obligatoirement une fin de ligne et rend <lb/> redondant. Par ailleurs, les limitations intrinsèques à la technologie XML Schema utilisée pour la rétroconversion du FEW ne permettent pas qu'apparaissent plus d'un <lb/> entre deux paragraphes dans le schéma FSML. Il est à remarquer que cette décision implique qu'on ne conserve pas cette partie de la mise en page du FEW papier, dans laquelle un espace blanc correspondant à une ligne peut apparaître entre deux paragraphes (la présence d'un <lb/> entre deux paragraphes aurait pu, sinon, servir à transcrire ce « blanc »).

2. Suppression des espaces excessifs éventuels autour des balises <lb/> ou <col/>.

3. Suppression des balises <lb/> qui se trouveraient juste avant une balise <col/>. En effet, la présence d'une balise <col/> implique obligatoirement une fin de ligne et rend <lb/> redondant.

Les deux dernières vérifications consistent en la détection de schémas suspects, menant à une demande de correction manuelle :

5. Demande de correction en cas de détection d'une balise fermante située juste après une balise <lb/> ou <col/>. En effet, les spécifications FFML demandent, en vertu de la logique interne du texte, mais aussi pour des raisons d'exploitation ultérieure (cf. ci-dessous), que les balises <lb/> soient les dernières balises de la ligne en cas de balises successives. Il est donc très probable que la balise fermante doit être déplacée avant la balise <lb/>.

6. Demande de correction en cas de présence de balises de formatage <i>, <b>, <sc> ou <e> autour d'un item <lb/> ou <col/> selon le schéma suivant : <i>texte </i> <lb/> <i>texte</i>. En effet, les spécifications FFML demandent que les segments de texte en italiques, grasses, petites capitales et exposants, s'ils sont coupés par une fin de ligne, ne soient pas répétés, mais balisés comme ceci : <i>texte <lb/> texte</i>.

Les deux dernières vérifications sont essentielles pour le bon fonctionnement des algorithmes de balisage. Les balises ajoutées par ces derniers risqueraient, sinon, d'être introduites à de mauvais endroits. Par exemple, l'algorithme *tag-form* se sert des balises <i>...</i> pour repérer un lexème ou une locution. Si un lexème ou une locution est coupé par une fin de ligne, comme casuel dans l'extrait suivant (FEW 2, 479b, CASUALIS I 1), et que le balisage soit

> [...] abearn. id., PtAud. <i>ca-</i><lb/>
> <i>suel</i>, Seudre S. „éventuel, probable", [...]

l'algorithme *tag-form* reconnaîtra et balisera deux lexèmes *ca-* et *suel* au lieu d'un seul lexème *casuel*.

5.3.2 Algorithmes de correction d'incohérences du FEW

5.3.2.1 Algorithme *streamline-layout-tags*

L'algorithme *streamline-layout-tags* concerne les balises de formatage (gras), <i> (italique) et <sc> (petites capitales). Il effectue des corrections nécessaires au bon fonctionnement de l'algorithme de balisage des étymons. En effet, la reconnaissance des étymons utilise comme indicateurs la présence de ces balises de formatage en plus de la présence de mots-clés provenant d'une liste des étymons du FEW ($\rightarrow$ 5.4.4). Ces mots-clés, comprenant par exemple "completus" ou "*abantiare", sont censés se trouver à l'intérieur des balises de formatage. Or, l'astérisque placé devant les étymons non attestés pose un problème particulier. N'étant pas nécessairement en grasses ou en italiques et jamais en petites capitales, il risque de ne pas être inclus, lors de la saisie FFML, à l'intérieur des balises de formatage. L'algorithme *streamline-layout-tags*, s'il trouve un astérisque placé juste devant une balise , <i> ou <sc>, le déplace après la balise de

formatage, afin par exemple que "*abantiare" devienne "*abantiare". Un avertissement est émis afin de signaler que le texte a été modifié.

Cet algorithme effectue encore une autre opération de nettoyage, consistant à supprimer les balises et qui se trouveraient autour d'un caractère isolé (éventuellement suivi d'un point). Cette mise en grasses est due à des erreurs de saisie ou de numérisation, ou pourrait également provenir du FEW lui-même. Les seuls cas autorisés, et donc non corrigés, sont des lettres isolées qui sont en grasses dans le FEW, car elles constituent des étymons-vedettes : 'A' (FEW 24, 1a), 'g' (FEW 4, 1a), 'i' (FEW 4, 530a), 'l' (FEW 5, 101a), 'o' (FEW 7, 260a), 'p' (FEW 7, 454a), 'r' (FEW 10, 1a), 's' (FEW 11, 1a), 'T' (FEW 13/1, 1a) et 'ō' (FEW 7, 260a). Mis à part le <A> majuscule et le <ō>, ces lettres sont directement suivies du point final de l'entrée, lui aussi en grasses. Les expressions régulières qui permettent de repérer les caractères *isolés* (l'isolement étant défini par la présence de caractères d'espacement ou de signes de ponctuation autour du caractère) – sauf ceux autorisés, cela s'entend – sont les suivantes :

- caractères isolés sauf <A> et <ō> :
  \s*[^Aō]\s*
- caractères isolés suivis d'un point, sauf <g>, <i>, <l>, <o>, <p>, <r>, <s> et <T> :
  \s*[^giloprsT]\.\s*

L'algorithme *streamline-layout-tags* cherche ces deux motifs à l'intérieur des éléments XML .... En cas de détection, il opère directement sur le texte en supprimant les balises de formatage en grasses et émet un avertissement afin de signaler la modification.

5.3.2.2 Algorithme *streamline-quotes*

L'algorithme *streamline-quotes* effectue plusieurs vérifications, suivies de corrections, concernant les signes de ponctuation du FEW qui vont par paires, c'est-à-dire les guillemets, les crochets et les semi-crochets (utilisés dans le FEW pour signaler une forme typisée). Les paires de parenthèses ne peuvent être vérifiées à ce stade, car le FEW se sert également des parenthèses fermantes pour les appels de note, qui sont balisées à un stade ultérieur par l'algorithme de balisage sémantique *tag-appelnote* (→ 5.4.2).

Deux opérations distinctes sont effectuées par l'algorithme *streamline-quotes*.

Une première opération consiste à vérifier la disposition des guillemets par rapport aux balises de formatage , <i>, <e> et <sc>. En effet, le corpus de test a montré plusieurs cas du type

<i> texte-en-italique „</i>texte-entre-guillemets“

où un guillemet ouvrant et une balise fermante (ici </i>, indiquant la fin d'un texte en italiques) ont été inversés. Il s'agit là du reflet d'erreurs typiques de numérisation, le guillemet ayant été transcrit en italique par l'OCR. L'algorithme *streamline-quotes* corrige ces erreurs en utilisant la propriété suivante : dans le FEW, les guillemets sont censés ne jamais être en italiques, en grasses, en exposants ou en petites capitales. L'algorithme consiste donc à itérer sur les éléments XML <i>...</i>, ..., <e>...</e> et

<sc>...</sc> et à vérifier qu'ils ne contiennent pas de guillemets. Dans le cas contraire, une alerte est émise, qui stoppe le traitement et invite à corriger le document FFML.

Une fois ce premier traitement effectué, l'algorithme *streamline-quotes* effectue un second traitement de l'article, consistant à vérifier que les guillemets, crochets et semi-crochets sont bien présents par paires et ne s'imbriquent pas. Cette vérification pose un problème particulier en ce qui concerne les guillemets. En effet, le FEW présente trois sortes de guillemets :

> “ (= left double quotation mark, code unicode U+201C)
> ” (= right double quotation mark, code unicode U+201D)
> „ (= double low-9 quotation mark, code unicode U+201E)

En général, le caractère U+201E fait dans le FEW office de guillemet ouvrant et le caractère U+201C de guillemet fermant, comme dans l'extrait suivant :

> Fr. *egre* adj. „acide au goût“ Chrestien [...] (FEW 24, 94b, ACER)

Toutefois, à partir de la page 689 du volume 25, ces règles ont été modifiées. Le caractère U+201C est devenu le guillemet ouvrant, tandis que c'est U+201D qui apparaît comme guillemet fermant :

> Neuch. *reaugment* m. “hausse (d'impôt)” (1676) [...] (FEW 25, 882b, AUGMENTUM)

La page 689 commence malheureusement en plein milieu d'un article (FEW 25, 687-691, ATRIUM), entre deux paragraphes. Dans un souci d'harmonisation interne, l'algorithme normalise les guillemets de la seconde partie de cet article en les remplaçant par leur correspondant dans la première partie.

L'algorithme de vérification de l'équilibrage des guillemets, crochets et semi-crochets se base ensuite sur la propriété suivante : à chaque article peuvent être associés, en fonction du volume (éventuellement du tome) et de la page où débute l'article, les guillemets ouvrant et fermant qu'il peut contenir et le guillemet illicite qu'il ne peut contenir.

Concrètement, l'algorithme itère sur tous les paragraphes de l'article. Pour chaque paragraphe, une chaîne virtuelle est créée, dans laquelle toutes les balises sont transparentes (excepté la balise </p>, qui termine la chaîne). L'algorithme itère sur tous les caractères de la chaîne et, lorsqu'il trouve un guillemet, crochet ou semi-crochet ouvrant licites, le mémorise. Lorsqu'il trouve un caractère fermant licite, il vérifie qu'il correspond au caractère ouvrant mémorisé. En cas de non-correspondance, une alerte est émise ; de même si deux caractères ouvrants, ou deux caractères fermants, se succèdent. L'algorithme ne tolère pas une suite de guillemets ou crochets récursifs du type "texte [texte[texte]]". En outre, une alerte est également émise si un guillemet illicite est détecté.

L'algorithme *streamline-quotes* est disponible en annexe (→ F.5).

5.3.3 Algorithme interprétant les tirets de fin de ligne : *merge-split-words*

Nous avons vu que la recherche de motifs requérait une stricte correspondance entre le texte du FEW et le motif recherché (→ 5.2.2). Les tirets de fin de ligne posent un problème à cet égard. En effet, le motif *Chambon* ne sera pas reconnu s'il se trouve à

1. Mfr. *substantif* adj. „qui forme la base de (p.
ex. de la vie religieuse, de l'amour, etc.)“ (14. jh.–
Desch, Froiss, Gdf; Lac); m. „résumé d'une narra-
tion, ne donnant que l'essentiel“ (1478, Leseur 1,
95, Bb).

FIGURE 5.1 – Exemple de tirets en fin de ligne (FEW 12, 357a, SUBSTANTIVUS)

cheval sur deux lignes dans la version papier du FEW, car un trait d'union vient alors s'insérer à l'endroit de la césure, comme ceci : *Cham-bon*. La solution consistant à mettre dans les listes de mots-clés, pour chacun d'eux, toutes les variantes possibles de coupures est fastidieuse et algorithmiquement peu optimale, car elle conduirait à augmenter de façon considérable la taille des listes. Une autre solution serait de supprimer du dictionnaire informatisé tous les traits d'union apparaissant à l'endroit d'une césure, qui n'ont d'utilité que dans la version imprimée du FEW. La question se pose dès lors de savoir comment les reconnaître et comment les supprimer.

5.3.3.1 Les tirets du FEW

Nous distinguons dans le FEW trois sortes de tirets :

- le *trait d'union* (-) est un petit trait horizontal joignant deux mots qui sont censés n'en faire qu'un (*Benoît de Sainte-Maure*) ;
- le *tiret de césure* (-) est un petit trait horizontal (formellement identique au trait d'union) apparaissant en fin de ligne lorsqu'un mot se poursuit à la ligne suivante (*Cham-bon*) ;
- le *vrai tiret* est un signe de ponctuation qui se présente sous deux formes, à savoir le tiret cadratin (—) ou le tiret semi-cadratin (–).

Il s'avère que ces trois types de tirets sont susceptibles d'apparaître en fin de ligne, y compris le vrai tiret (cf. figure 5.1).

Les spécifications FFML ($\rightarrow$ 4.3) réduisent les possibilités de codage de ces tirets à trois caractères Unicode : U+002D (trait d'union ou tiret de césure), U+2014 (tiret cadratin) et U+2013 (tiret semi-cadratin). En théorie et comme le montre l'extrait ci-dessus, le tiret semi-cadratin ne se confond pas avec les autres tirets. En pratique toutefois, la confusion occasionnelle entre le caractère U+2013 (tiret semi-cadratin) et le caractère U+002D (trait d'union ou tiret de césure) est inévitable, soit par erreur du FEW lui-même, soit lors de la saisie ou de la numérisation du texte.

On ne peut donc utiliser ni la position du tiret en fin de ligne ni son codage comme critères de détection du tiret de césure. Les caractères U+002D apparaissant en fin de ligne ne peuvent être purement et simplement supprimés, sous peine de perdre des tirets sémantiquement pertinents.

5.3.3.2 L'attribut *merge-split-words*

Distinguer entre les différentes sortes de tirets présente un intérêt uniquement lorsqu'un tiret de césure coupe un motif et empêche de ce fait la reconnaissance de ce dernier par

les algorithmes. Afin d'éviter tout risque de perte d'information en cas de mauvaise interprétation du FEW, nous décidons de ne supprimer aucun tiret, mais d'ajouter à chaque balise <lb/>, équivalent au passage à la ligne suivante, un attribut indiquant si les items qui entourent la balise constituent ou non un ensemble à fusionner lors de la recherche de motifs. Cet attribut servira d'indicateur pour les algorithmes de balisage : si sa valeur indique que les items situés de part et d'autre de la balise <lb/> forment un ensemble joint, la recherche de motifs s'appliquera au texte du FEW comme si la balise <lb/> et le tiret de césure qui la précède étaient invisibles. Cette représentation permet de conserver dans le texte les tirets d'origine, tout en donnant une information sur la façon dont il faut les interpréter lorsque cela est utile en pratique.

L'attribut *merge-split-words* appliqué à chaque balise <lb/> reçoit une des trois valeurs suivantes :

1. "no" si les deux parties de texte situées respectivement avant et après <lb/> ne doivent pas être fusionnées. C'est le cas lorsqu'il s'agit d'un vrai tiret, par exemple dans une fourchette comme "BeaumCout-<lb/>Trév 1752" :

    mfr. id (1577) ; fr. „tout ce qui se rapporte au
    pouvoir, aux intérêts terrestres“ (BeaumCout-
    Trév 1752, Li [...] (FEW 13/1, 181b, TEMPORALIS II 1)

2. "keep-dash" si les deux items doivent être fusionnés en conservant le tiret. C'est le cas lorsqu'il s'agit d'un trait d'union, comme dans "St-Amand", "St-Léger", "Benoît de Sainte-Maure", "AnnChâteau-Th." etc.

3. "discard-dash" si les deux items doivent être fusionnés sans le tiret. C'est le cas lorsqu'il s'agit d'un tiret de césure, par exemple dans "GuernesS-<lb/>Thomas" (FEW 8, 273a, PERSONALIS).

5.3.3.3 Méthode d'identification des tirets

La solution choisie pour déterminer la valeur de l'attribut *merge-split-words* est très pratique. Elle se base sur les listes de mots-clés, qui constituent un ensemble de lexiques ventilés par domaines notionnels de la langue fewienne (→ 5.2.2.1).

Si la fusion des deux bouts de texte, le tiret ayant été enlevé, constitue un mot-clé, on peut légitimement considérer que le tiret représente un tiret de césure. Toutefois, les listes de mots-clés ne sont actuellement pas toutes exhaustives ; par ailleurs, elles ne couvrent pas tous les types de mots du FEW susceptibles d'être coupés par une fin de ligne. L'application de cette seule règle laisserait donc comme vrais tirets de nombreux tirets de césure, notamment au sein du commentaire.

Pour limiter les erreurs, nous identifions également les vrais tirets. Pour ce faire, nous nous servons encore des listes de mots-clés. Si l'un des deux items situés de part et d'autre de la balise <lb/> constitue un mot-clé, il est en effet fort probable que le tiret soit un vrai tiret, séparant par exemple deux éléments d'une fourchette de datation. La même conclusion s'impose si l'une des deux parties est composée de chiffres.

Dès lors, l'attribution d'une valeur à l'attribut *merge-split-words* s'effectue en quatre temps :

1. Si la fusion des deux items, le tiret ayant été conservé, forme un mot-clé, l'attribut reçoit pour valeur "keep-dash" (trait d'union).

2. Si la fusion des deux items, le tiret ayant été supprimé, forme un mot-clé, la valeur de l'attribut est "discard-dash" (tiret de césure).

3. Si l'un des deux items est un mot-clé ou est composé de chiffres, la valeur de l'attribut est "no" (vrai tiret).

4. Si aucune des trois propriétés précédentes n'est vérifiée, le tiret est considéré par défaut comme un tiret de césure : la valeur est "discard-dash".

L'ordre de vérification de ces propriétés est important. En effet, il est possible qu'un mot-clé soit identique à la suite de deux autres mots-clés. Par exemple, "BeaumCout", sigle bibliographique qui renvoie aux *Coutumes du Beauvaisis* de Philippe de Beaumanoir, un texte d'ancien français du 13e siècle, équivaut par hasard à la suite *Beaum* (présentation abrégée des *Œuvres poétiques* du père de l'auteur des *Coutumes*) et *Cout* (abréviation d'un manuel de vénerie français de 1890 intitulé *Le Couteulx de Canteleu*). Le fait de d'abord considérer l'ensemble fusionné avant de considérer les parties permet de détecter et de résoudre les quelques cas ambigus de ce genre.

Avant de vérifier les trois propriétés énumérées ci-dessus, il est en outre nécessaire de vérifier que la balise <lb/> est bien entourée de deux morceaux de texte et que celui situé à sa gauche se termine bien par un tiret, sans quoi la valeur de l'attribut sera d'office "no". Remarquons qu'il n'y aura jamais de balise devant <lb/> (en cas de mots en italique par exemple) puisqu'elle aura été supprimée par l'algorithme *streamline-breaks* (qui doit donc obligatoirement intervenir avant *merge-split-words* dans la séquence). Enfin, si le tiret de fin de ligne est précédé d'un délimiteur, c'est-à-dire d'un caractère qui ne peut faire partie d'un mot (une liste de délimiteurs a été constituée spécifiquement pour le FEW, → 4.2.2), le texte ne peut être fusionné : la valeur de l'attribut sera "no" également.

La description formelle de l'algorithme *merge-split-words* est disponible en annexe (→ F.2).

5.3.3.4 Remarque

Il est important de remarquer que l'algorithme *merge-split-words*, parce qu'il se base sur des listes de mots-clés et non sur une analyse fine du texte, ne conduit pas nécessairement à une interprétation "juste" linguistiquement parlant. Nous ne pouvons assurer que les tirets seront toujours analysés de façon correcte par rapport à leur rôle exact dans le FEW. Cependant, l'algorithme remplira systématiquement son objectif, qui est de permettre la détection des mots-clés provenant des listes, puisqu'il se sert de ces mêmes mots-clés pour déterminer la valeur du tiret et que les listes de mots-clés ne sont pas modifiées entre l'application de *merge-split-words* et l'application des algorithmes de balisage.

L'objectif très pratique de *merge-split-words* explique pourquoi cet algorithme est classé dans les algorithmes de prétraitement et non dans les algorithmes de balisage sémantique : son but n'est en aucun cas de donner une information fiable sur le sens des tirets du FEW, mais simplement de permettre l'application des algorithmes suivants. Cette réserve mise à part, l'attribut donné à la balise <lb/> pourrait se révéler utile pour la visualisation d'un article sur écran si l'on décidait de ne pas conserver la mise en page initiale du FEW papier. L'utilisation de cet attribut dans ce but de visualisation ne

pourrait toutefois avoir lieu qu'après la rétroconversion de tous les articles du FEW, de façon à s'assurer que les listes de mots-clés sont complètes (→ 6.6).

5.4 Algorithmes de balisage

5.4.1 Balisage des notes : tag-notes

<notes> 1) Nicht <i>ablativo</i>, wie SainéanEt 2, 359 schreibt. </notes>	<notes count="1"> <note id="1">1) Nicht <i>ablativo</i>, wie SainéanEt 2, 359 schreibt.</note> </notes>

(FEW 24, 34 ABLATIVUS)

5.4.1.1 Objectifs

L'algorithme *tag-notes* a pour objectif de reconnaître et de baliser les notes de fin d'article, regroupées dans le champ des notes (→ 3.7.1). Ce balisage se justifie non seulement pour la lecture du FEW dans sa dimension monographique (balisage combiné des notes et appels de note, → 3.7.4), mais aussi pour la bonne application des algorithmes *tag-appelnote*, *split-doc-com* et *tag-signature* (voir ces algorithmes).

Chaque note constitue dans le FEW un paragraphe distinct, balisé <p>...</p> dans le document FFML (→ 4.3.2). Conformément à la modélisation proposée (→ 3.7.4), l'algorithme *tag-notes* doit baliser chacune des notes qui se trouve dans le champ des notes en insérant une balise <note> à l'intérieur de chaque balise <p>.

Tag-notes a également pour tâches d'expliciter le numéro de la note, au moyen d'un attribut *id* associé à la balise <note>, et d'ajouter à la balise <notes> indiquant le début du champ des notes un attribut *count* ayant pour valeur le nombre total de notes de l'article. Cet attribut doit permettre aux algorithmes suivants de connaître le nombre de notes de l'article sans devoir les recompter.

Le résultat attendu de l'application de l'algorithme *tag-notes* sur un article contenant deux notes est donc le suivant :

```
<notes count="2">
<p><note id="1">1) Ici se trouve la première note.</note></p>
<p><note id="2">2) Ici se trouve la seconde note.</note></p>
</notes>
```

nebeneinander; ersteres scheint zu anfang des 17. jhs. im begriff gewesen zu sein, zu verschwinden (Mon 1636: ,vieux'). — II 1 ist eine latinisierende redensart, 2 aus GEMITUS „das stöhnen" entlehnt. S. noch *GÈMICARE. — ML 3722; Risop.

1) Ebenso piem. *gem* „gemito", Castellinaldo *džam* AGl 16, 521, ven. *zémo* SFR 7, 221, Disentis *žęm*.
2) Hier offenbar mit JEHAN gekreuzt.

FIGURE 5.2 – Exemple de notes (FEW 4, 93a, GEMERE)

5.4.1.2 Reconnaissance du champ des notes

Indicateurs de reconnaissance

> IPo. Le champ des notes se trouve tout à la fin de l'article. Il est systématiquement précédé d'une ligne de séparation "qui s'étend environ jusqu'au quart de la colonne" (Büchi 1996, 137).

Les possibilités de détection du champ des notes diffèrent selon le mode d'acquisition du texte fewien.

En cas de saisie manuelle du FEW (mode d'acquisition considéré comme acquis lors de la définition du format FFML, → 4.2.1), le champ des notes est censé avoir été reconnu et balisé manuellement. Ce balisage manuel se fonde sur la ligne de séparation entre le commentaire et le champ des notes (cf. Büchi 1996, 137), ligne qui constitue un moyen de reconnaissance fiable dans la version imprimée du FEW, même par un humain non initié aux structures de l'ouvrage (cf. figure 5.2).

Conformément aux directives FFML (→ 4.3.2), des balises <notes>...</notes> sont dès lors insérées autour du champ des notes dans les documents FFML avant rétroconversion. L'algorithme *tag-notes* ne doit donc pas générer ce balisage ; au contraire, il peut s'appuyer sur le balisage préalable du champ des notes pour détecter si l'article comporte des notes à traiter.

En cas d'impossibilité de saisir manuellement le FEW, il est nécessaire de trouver un moyen de balisage automatique du champ des notes. La ligne de séparation entre le commentaire et le champ des notes n'est pas reconnue systématiquement par les logiciels d'océrisation et ne peut donc être utilisée comme critère de détection. Le seul moyen de baliser le champ des notes de façon automatique consiste à utiliser le balisage préalable des notes individuelles (→ 5.4.1.3). Une fois que ces dernières sont balisées, il est en effet possible de les rassembler en un seul élément XML <notes>.

Traitement des incohérences et erreurs

Les articles de la partie des Inconnus (volumes 21-23) ne constituent pas un cas particulier. À condition de considérer qu'en structure de surface, un article des volumes 21-23 correspond aux lexèmes regroupés sous un concept (→ 3.5.2), on peut affirmer que les notes apparaissent dans ces volumes de la même façon que dans les matériaux

étymologisés.

> Le FEW présente un cas, qui n'est peut-être pas unique, de notes ne se situant pas en fin d'article mais en fin de colonnes (FEW 24, 444a-b).

Ce cas atypique ne peut être résolu qu'en le normalisant, solution qui a pour inconvénient de ne pas rester fidèle à la mise en page de la version papier du FEW, mais qui se justifie amplement dans une version informatisée. Les notes devront être considérées dans le document FFML comme apparaissant à la fin de l'article (donc p. 444b), suivant la règle générale.

5.4.1.3 Reconnaissance des notes individuelles

Indicateurs de reconnaissance

> IPo. Les notes sont situées en fin d'article, dans le champ des notes. Chacune d'elles constitue un et un seul paragraphe.
> ITe. Chacune des notes commence par un nombre en chiffres arabes suivi d'une parenthèse fermante.

Il est possible de détecter un paragraphe de notes en vérifiant qu'il commence par un ou plusieurs chiffre(s) arabe(s) suivi(s) d'une parenthèse fermante. Cette vérification peut être faite au moyen d'une expression régulière telle que [1-9][0-9]*\), désignant tout nombre à partir de 1[6].

> On sait en outre que la première des notes débute toujours (sauf erreurs) par le chiffre 1. Les notes se succèdent ensuite en ordre croissant, de 1 à n.

On peut vérifier en même temps l'ordre croissant (ou décroissant, selon le sens du parcours) des numéros relevés par l'expression régulière. Le nombre détecté par l'expression régulière dans le dernier paragraphe de l'article est censé donner le nombre total de notes, qui peut être confirmé par le nombre total de paragraphes de notes reconnus.

Le sens du parcours dépend des conditions d'acquisition du texte fewien. En cas de saisie manuelle, des balises <notes>...</notes> délimitent le champ des notes (→ 4.3.2). Le parcours peut dès lors s'effectuer à partir du premier paragraphe après la balise <notes> :

```
|[
VAR
    n : integer ;
    ck : chunk
BEGIN
"initialiser ck à la balise ouvrante <notes>"
```

[6]Nous n'intégrons pas dans les expressions régulières ici exposées les caractères d'espacement qui précèdent et/ou suivent éventuellement l'expression. Ces caractères d'espacement (\s*) sont ajoutés aux expressions régulières dans le métalangage et sont bien sûr pris en compte dans l'implémentation.

```
; IF ck = null →
    SKIP
[] ck != null →
    n := 0
    ; "déplacer ck à la première balise ouvrante <p> après <notes>"
    ; DO ck != null →
        n := n + 1
        ; "traiter la note"
        ; "vérifier que le numéro de note est identique à n"
        ; "déplacer ck à la balise ouvrante <p> suivante"
    OD
    ; "déplacer ck à la balise ouvrante <notes>"
    ; "ajouter à <notes> un attribut count de valeur n"
FI
END
]|
```

En cas d'impossibilité de saisie manuelle du champ des notes, l'absence de balisage préalable du champ des notes ne permet pas de connaître l'endroit où se situerait la première note de l'article. Le parcours doit dès lors s'effectuer à partir de la fin de l'article, donc de la dernière note. La détection de l'expression régulière dans ce dernier paragraphe est l'indice qu'il existe des notes à traiter. Le balisage de chaque note s'effectue paragraphe par paragraphe en remontant jusqu'à la première des notes, reconnaissable par la détection du numéro "1" dans le résultat de l'expression régulière :

```
|[
VAR
    n : integer ;
    ck : chunk
BEGIN
; "initialiser ck à la dernière balise ouvrante <p> avant </art>"
; "vérifier qu'il s'agit d'une note"
; IF "c'est une note" →
    n := "récupérer le numéro de note"
    ; n_total := n
    ; DO n >= 1 →
        "traiter la note"
        ; "vérifier que le numéro de note est identique à n"
        ; "déplacer ck à la balise ouvrante <p> précédente"
        ; n := n - 1
    OD
    ; "déplacer ck à la balise ouvrante <notes>"
    ; "ajouter à <notes> un attribut count de valeur de n_total"
[] "ce n'est pas une note" →
    SKIP
FI
END
]|
```

Traitement des incohérences et erreurs

> Il se peut que la succession des numéros de note ne soit pas strictement croissante, par exemple en cas de saut d'un numéro ou de répétition d'un même numéro. Ces erreurs peuvent provenir du FEW ou d'une mauvaise saisie des documents FFML.

L'algorithme vérifie le bon ordonnancement des numéros de note en comparant chaque numéro de note avec le numéro du paragraphe dans lequel il se trouve (numéro donné par un compteur de paragraphes (n), incrémenté ou décrémenté selon le sens du parcours). En cas de non correspondance, un avertissement est émis pour inspection par un expert humain.

5.4.1.4 Résumé

Dépendances

La reconnaissance des notes nécessite le balisage préalable de l'article, des paragraphes et du champ des notes, tous trois prévus dans le format FFML. En cas de non saisie manuelle des articles du FEW, l'impossibilité de disposer du balisage préalable du champ des notes n'est pas problématique, mais conduit à certaines modifications algorithmiques, notamment concernant le sens du parcours (cf. ci-dessous).

Description de l'algorithme en cas de saisie manuelle du FEW

Tag-notes a pour objectif de baliser les notes individuelles dans les articles qui contiennent un champ de notes. Une première condition vérifie la présence de la balise <notes>. Une boucle est ensuite opérée sur les paragraphes inclus dans <notes>...</notes>, en incrémentant un compteur de paragraphes.

Pour chaque paragraphe, une chaîne virtuelle est créée dans laquelle toutes les balises sont rendues transparentes, sauf </p> qui clôture la chaîne. L'algorithme cherche au début de cette chaîne l'expression régulière suivante : /[1-9][0-9]*\)/.

L'algorithme détecte une erreur et émet un avertissement si l'expression régulière n'a pas été trouvée ou s'il n'y a pas de correspondance entre le numéro du paragraphe (donné par le compteur de paragraphes) et le numéro de note détecté par l'expression régulière.

Si le numéro trouvé par l'expression régulière est en accord avec le numéro du paragraphe, une balise <note> est insérée au tout début de la chaîne, juste après la balise <p>. La balise reçoit un attribut *id* ayant pour valeur ce numéro. Une balise fermante </note> est ensuite insérée à la fin du paragraphe, juste avant la balise </p>.

Une fois que la boucle est terminée et que toutes les notes ont été balisées, un attribut *count* ayant pour valeur le nombre total de notes trouvées est ajouté à la balise <notes> qui marque le début du champ des notes.

Description de l'algorithme en cas de non saisie manuelle du FEW

En cas d'impossibilité de saisie manuelle, l'algorithme *tag-notes* présenté ci-dessus doit être modifié de façon à ne pas dépendre de l'élément <notes>. C'est la détection de l'expression régulière dans le dernier paragraphe de l'article qui signale la présence de notes à traiter. Le nombre détecté par l'expression régulière dans ce dernier paragraphe indique le nombre de notes. Le parcours a lieu dans un sens décroissant jusqu'à la détection de la première note, signalée par la détection du chiffre 1 dans l'expression régulière /[1-9][0-9]*\)/. Après le balisage de toutes les notes individuelles, le balisage du champ des notes a lieu par insertion des balises <notes> et </notes> respectivement au début et à la fin de l'ensemble des notes.

5.4.2 Balisage des appels de note : tag-appelnote

mars. <i>gravo</i> „marche d'escalier" <e>2</e>) M	mars. <i>gravo</i> „marche d'escalier" <appelnote id="2" status="ok"><e>2)</e></appelnote> M

(FEW 4, 204b, GRADUS)

5.4.2.1 Objectifs de l'algorithme

L'algorithme *tag-appelnote* a pour objectif de baliser tous les appels de note présents dans un article et d'identifier pour chacun le numéro de la note.

Ce balisage a pour but premier de faciliter la navigation et la lecture des articles longs en permettant que chaque appel de note puisse être relié a posteriori (lors de la future phase d'exploitation du dictionnaire rétroconverti) à la note correspondante (→ 3.7.4). Par ailleurs, leur balisage présente un intérêt indirect pour deux raisons. Tout d'abord, certains algorithmes, qui balisent des éléments situés entre parenthèses (cf. *tag-precisions*), utilisent ces parenthèses comme critères de détection. La parenthèse qui suit un appel de note risque d'être interprétée par ces algorithmes comme une "vraie" parenthèse, c'est-à-dire comme une parenthèse fermant une parenthèse précédemment ouverte. La détection préalable des appels de note permet d'éviter ce risque de confusion. Une seconde raison est la nécessité de discriminer les appels de note et les numéros d'édition, qui sont eux aussi en exposant et qui peuvent se trouver par hasard devant une parenthèse fermante, par exemple dans (Corom^2), où *Corom*^{2} désigne la deuxième édition du dictionnaire étymologique espagnol de Corominas.

L'élément XML <appelnote> doit inclure la parenthèse fermante qui suit le numéro d'appel de note. Ce dernier doit être repris comme valeur d'un attribut *id*, comme ceci :

texte<appelnote id="1">1)</appelnote> texte

L'algorithme doit également identifier et signaler des problèmes éventuels de non-correspondance entre les notes et les appels de note, au moyen d'un attribut (*call-sequence*) ajouté à la balise <notes> qui indique le début du champ des notes.

2. Fr. *génitif* „cas auquel sont mis, dans les langues à déclinaisons, les compléments de nom et certains régimes“ (dp. 14e s.), *cas génitif* Ch d'Orléans[1]), *génitif* „rapport de possession marqué par la préposition de“ (Nic 1606—Trév 1704)[2]).

FIGURE 5.3 – Exemple d'appels de notes (FEW 4, 102b, GENITIVUS)

La valeur de cet attribut doit indiquer si le nombre de notes et le nombre d'appels de note sont égaux et si les appels de note apparaissent dans le texte en suivant un ordre croissant.

Quatre valeurs sont possibles pour l'attribut *call-sequence* :

– "ok", indiquant une correspondance exacte entre les numéros des notes et les numéros des appels de note. Cela signifie que tous les appels de note ont été trouvés par l'algorithme et qu'ils se présentent dans le texte en ordre croissant (de 1 à N, N représentant le nombre de notes de l'article, c'est-à-dire la valeur de l'attribut *count* de <notes>) ;
– "incomplete", indiquant que tous les appels de note n'ont pas été retrouvés par l'algorithme, soit par échec de l'algorithme, soit par erreur du FEW ou du travail de saisie ;
– "non-monotonic", indiquant que tous les appels de note ont été trouvés, mais qu'ils n'apparaissent pas dans le texte en ordre croissant ;
– "some-redundancy", indiquant que tous les appels de note ont été trouvés, mais que certains se répètent, c'est-à-dire qu'à une note correspond plus d'un appel de note (soit par échec de l'algorithme, soit par erreur de la saisie, soit parce que le FEW fait effectivement appel à plusieurs reprises à une même note, ce qui est en accord avec ses principes lexicographiques).

5.4.2.2 Détection des appels de note

Détection d'un candidat appel de note

> IPo. Les appels de note peuvent apparaître à n'importe quel endroit de l'article, excepté dans le champ des notes.
> ITe/ITy. Ils sont composés de chiffres arabes en exposant, suivis d'une parenthèse fermante non mise en exposant (cf. figure 5.3).

Suivant les indicateurs textuels, un appel de note pourrait être reconnu au moyen d'une expression régulière telle que <e>[1-9][0-9]*</e>\) ou, si les balises XML ne sont pas intégrées dans l'expression régulière, par [1-9][0-9]*\) combiné à l'obligation que des balises <e>...</e> soient présentes autour du numéro en chiffres arabes. Les critères structurels excluent le champ des notes du texte dans lequel doit se faire la recherche.

Ces critères sont fiables : la présence d'un chiffre en exposant suivi d'une parenthèse fermante s'applique à tous les appels de notes et ne devrait donc pas générer

de faux négatifs (→ 5.2.1). Elle n'est cependant pas assez spécifique pour identifier de façon certaine un appel de note, puisqu'elle donnerait lieu à de faux positifs (→ 5.2.1) :

> Certains éléments ressemblent à des appels de note mais n'en sont pas. Par exemple, les numéros d'édition sont en exposant dans le FEW et pourraient se trouver par hasard à la fin d'un extrait de texte situé entre parenthèses, par exemple dans (*AlcM ; Corom*^{2}).

Discriminer les appels de note et les numéros d'édition en balisant préalablement les références bibliographiques est une solution peu efficace, car tous les sigles bibliographiques ne peuvent pas être détectés avec certitude (cf. *tag-biblio*). Une solution plus sûre pour lever ces ambiguïtés consiste à examiner les parenthèses qui se trouvent dans les environs proches du candidat appel de note.

Examen des successions de parenthèses

Appelons X le numéro en exposant susceptible d'être un appel de note. La détection de l'expression régulière [1-9][0-9]*\) assure qu'on a trouvé "X)". L'étape suivante consiste à déterminer si la parenthèse qui suit X est une parenthèse d'appel de note ou une "vraie" parenthèse. Pour ce faire, l'algorithme doit examiner si une parenthèse ouvrante ne se trouve pas à gauche de X. Le texte qui précède X (contexte gauche) doit être examiné de façon très large, depuis le début du paragraphe. L'algorithme recherche la parenthèse qui se trouve la plus proche de X à sa gauche. Quatre scénarios sont possibles :

1. il n'y a pas de contexte gauche ;
2. le contexte gauche ne contient aucune parenthèse : "...X)" ;
3. le contexte gauche contient une parenthèse fermante : ")...X)" ;
4. le contexte gauche contient une parenthèse ouvrante : "(...X)".

Le premier scénario est pour le moins étrange et méritera d'être signalé par une alerte. Le deuxième et le troisième laissent déduire que X est très certainement un appel de note, que nous pouvons baliser comme tel. Le quatrième scénario est problématique et requiert un examen plus approfondi. Un cas particulier doit en effet être pris en compte :

> IPo. Un appel de note peut apparaître au milieu d'un extrait de texte situé
> entre parenthèses :
> <i>macédoine littéraire</i> „assem-<lb/>
> blage de pièces de différents genres dans un<lb/>
> même livre“ (1786 <e>4</e>), Bachaumont, s. Li ; ‚péjora-<lb/>
> tif' Boiste 1829 ; s. noch Lar 1873), <i>macédoine</i><lb/>
> (FEW 6/1, 3b, MACEDONIA)

Pour savoir si X est un appel de note dans l'expression "(...X)", il nous faut examiner le texte qui suit X (contexte droit), afin de voir s'il se trouve à droite de X une autre parenthèse fermante.

L'examen du contexte droit mène à quatre scénarios :

1. il n'y a pas de contexte droit : "(...X)"
2. le contexte droit ne contient aucune parenthèse : "(...X)..."
3. le contexte droit contient une parenthèse ouvrante : "(...X)...("
4. le contexte droit contient une parenthèse fermante : "(...X)...)"

Les trois premiers scénarios permettent d'induire que X n'est pas un appel de note, puisque la parenthèse ferme une parenthèse précédemment ouverte. Dans le quatrième cas, il s'agit très certainement d'un appel de note. Il reste à vérifier qu'un chiffre en exposant ne se trouve pas devant la deuxième parenthèse fermante, ce qui amènerait à une situation ambiguë du type "(...X)...Y)" (avec Y représentant un chiffre en exposant), qui ne pourrait être résolue que par un expert linguiste. S'il n'y a pas d'exposant devant la deuxième parenthèse fermante, X est bien un appel de note que nous pouvons baliser.

Remarques.

1. Il est important de souligner que cette méthode de vérification des parenthèses ne fonctionne que si on respecte l'invariant "ce qui se trouve avant X a déjà été traité, et les appels de note y ont été reconnus". La détection des appels de note doit donc absolument se faire linéairement du début vers la fin du paragraphe.
2. On pourrait aller plus loin encore dans le nombre de parenthèses à prendre en compte, pour envisager des cas comme "(...X)...Y)...)" ou "(...X)...(...)...)". Il nous semble cependant que cela reviendrait à complexifier l'algorithme de façon trop importante par rapport au petit nombre de tels cas particuliers qui pourraient se trouver dans le FEW. Une vérification de la correspondance entre les notes et les appels de note ($\rightarrow$ 5.4.2.3) constitue déjà une bonne alternative pour détecter ces cas particulièrement complexes et les baliser ensuite de façon manuelle.
3. Il est à remarquer également que cet examen de la succession des parenthèses n'est pas fiable à cent pour cent, car il ne prend pas en compte la possibilité que des parenthèses soient manquantes ou redondantes, par erreur du FEW ou par erreur de saisie. Malheureusement, la détection des appels de note est un préalable à la vérification des paires de parenthèses (cf. l'algorithme de prétraitement *streamline-quotes*, qui vérifie les paires de guillemets mais pas les parenthèses). L'algorithme *tag-appelnote* risque donc d'échouer en cas de parenthèses qui seraient manquantes dans le FEW ou auraient été oubliées lors de la saisie. La probabilité (faible) que de telles erreurs apparaissent justement là où se trouve un appel de note existe, même si aucun problème de ce genre n'a été relevé dans le corpus de test. De nouveau, une vérification de la correspondance entre les notes et les appels de note ($\rightarrow$ 5.4.2.3) constitue une solution permettant de détecter ces problèmes.

Acceptabilité d'un numéro d'appel de note

On pourrait observer que dans le scénario "(...X)...Y)" mentionné ci-dessus, l'ambiguïté pourrait être levée de façon automatique par l'examen des numéros X et Y. En

effet, si (au moins) l'un des deux nombres est supérieur au nombre total de notes de l'article, il ne peut s'agir d'un appel de note. La remarque vaut également pour les scénarios précédents où seul X est en cause : un appel de note "8)" trouvé dans un article qui ne contient que quatre notes pourrait être évincé sans qu'il soit besoin de vérifier les successions de parenthèses. Vérifier que X est licite, c'est-à-dire qu'il est inférieur ou égal au nombre total de notes de l'article (préalablement calculé par *tag-notes* et disponible dans l'attribut *count* de <notes>), est un moyen très simple de discriminer les vrais et les faux appels de note, en tous cas de détecter un certain nombre de faux appels de note.

Dans la pratique, ce critère n'est toutefois pas totalement fiable. Le corpus de test a montré des erreurs de saisie, où un appel de note "3)" avait été transcrit en "8)". Le non balisage de cet appel de note aurait empêché la détection de l'erreur et sa correction. Nous avons donc décidé de vérifier l'acceptabilité des numéros d'appel de note trouvés, mais sans les évincer pour autant. L'acceptabilité de chaque appel de note détecté est mémorisée dans une variable booléenne (*status_note_id*) qui reçoit la valeur "false" si le numéro détecté est supérieur au nombre total de notes et "true" dans les autres cas. Cette variable est utilisée lors du balisage de l'appel de note : un avertissement est émis en cas de valeur "false" pour vérification par un linguiste.

Incohérences

> ITy. Dans la refonte au premier volume du FEW (volumes 24 et 25 et articles publiés sur Internet), la parenthèse fermante qui suit le numéro d'appel de note peut se trouver en exposant (voir par exemple FEW 25, 882b, AUGMENTUM).

Dans les articles de la refonte présentant les parenthèses d'appel de note en exposant, il n'y a aucune ambiguïté : la parenthèse en exposant ne peut jamais fermer une parenthèse précédemment ouverte. La vérification des successions de parenthèses n'est pas nécessaire. Ce cas particulier sera donc traité dès le début de l'algorithme : si la parenthèse est en exposant, l'appel de note sera balisé sans examen des successions de parenthèses.

5.4.2.3 Correspondance des notes et des appels de note

Problématique

L'examen de la succession des parenthèses n'est pas totalement fiable : il peut échouer à cause de cas complexes non prévus dans l'algorithme ou à cause d'erreurs typographiques du FEW (parenthèses manquantes ou redondantes). Une vérification de la correspondance entre les notes et les appels de note apparaît comme un critère intéressant pour détecter ces problèmes, à condition que cette correspondance soit effectivement

la règle dans les articles du FEW.

> En général, les appels de note apparaissent dans l'ordre croissant, mais des incohérences sont possibles. Dans l'article DRILLEN (FEW 3, 159ab) par exemple, qui contient quatre notes, les appels de note apparaissent dans l'ordre 2-1-4-3.
> Par ailleurs, une même note peut être appelée à plusieurs reprises (cf. FEW 4, 584b, IMPERATIVUS). Ce cas est tout à fait permis dans le discours fewien.

Les causes d'une non correspondance exacte entre les notes et les appels de note peuvent être diverses. La redondance d'un même appel de note, par exemple, peut provenir d'un échec de l'algorithme, mais elle peut aussi être voulue par le rédacteur du FEW. Enfin, elle peut aussi résulter d'une erreur de saisie, qu'il serait intéressant de détecter, car elle peut alors être combinée à un appel de note manquant. C'est le cas si un appel de note, par exemple "3", a été pris pour un autre, par exemple "8", lors de la saisie ou de la numérisation d'un article (en l'occurrence, comprenant plus de sept notes) : il manque alors l'appel de note "3", tandis que l'appel de note "8" est redondant.

Ces cas de figure montrent qu'il serait vain de vouloir utiliser le nombre d'appels de note ou leur ordre croissant comme critères décisifs de reconnaissance : l'algorithme échouerait partout où les appels de note se présentent de façon non attendue. Nous décidons donc de vérifier les propriétés de correspondance entre les notes et les appels de note, sans toutefois utiliser cette vérification comme un critère de décision. La non correspondance éventuelle est signalée dans un attribut de la balise <notes> (indiquant le début du champ des notes) et fait l'objet d'un avertissement pour relecture par un expert humain.

Séquençage des appels de note

L'algorithme ajoute donc à la balise <notes> un attribut *call-sequence* qui détermine si la correspondance entre la séquence d'appels de note et la séquence de notes est correcte ou pas. Plusieurs cas de figure peuvent se présenter.

– Correspondance exacte entre les numéros des notes et les numéros des appels de note. Cela signifie que tous les appels de note ont été trouvés par l'algorithme et qu'ils se présentent dans le texte en ordre croissant (de 1 à N, N représentant le nombre de notes de l'article, c'est-à-dire la valeur de l'attribut *count* de <notes>). La valeur sera alors "ok".

– Tous les appels de note n'ont pas été trouvés par l'algorithme. La valeur sera alors "incomplete".

– Les appels de note n'apparaissent pas dans le texte en ordre croissant. La valeur sera alors "non-monotonic".

– À une note correspond plus d'un appel de note. La valeur sera alors "some-redundancy".

L'attribut a donc quatre valeurs possibles et exclusives. Une priorité a été définie pour les situations où plusieurs valeurs sont applicables : "incomplete" est prioritaire sur "non-monotonic", lui-même prioritaire sur "some-redundancy". L'absence d'un appel de note est en effet toujours problématique, alors que la redondance peut être due à des causes diverses. Dans le cas de figure où un appel de note est manquant et un autre redondant, la valeur définie sera "incomplete" et non "some-redundancy", puisque "incomplete" est prioritaire. Cet ordre de priorité explique la sémantique donnée aux quatre valeurs dans les objectifs de l'algorithme (→ 5.4.2.1).

Algorithme de vérification de la correspondance entre notes et appels de note

Afin d'examiner la correspondance entre les notes et les appels de note, une suite d'instructions spécifique ("vérifier correspondance", → 5.4.2.3) est appliquée à la fin de l'algorithme, après détection de tous les appels de note. Les numéros de note détectés ainsi que leur acceptabilité (→ 5.4.2.2) ont été mémorisés dans une structure de données adéquate au fur et à mesure du balisage des appels de note et peuvent dès lors être passés en revue.

Pour chaque appel de note considéré comme acceptable, l'algorithme vérifie que son numéro est supérieur à celui de l'appel de note précédent. L'information est mémorisée dans une variable booléenne (*ordre_croissant*). Un deuxième parcours de la séquence d'appels de note permet ensuite de mémoriser le nombre d'appels de note redondants (*repeated_notes*) ainsi que le nombre d'appels de note manquants (*missing_notes*).

Le résultat de ces deux parcours conduit à définir une valeur, de la façon suivante :

- Si certains appels de note n'ont pas été retrouvés, la valeur est définie comme "incomplete" ;
- Si les numéros ne se suivent pas en ordre croissant, la valeur est "non-monotonic" ;
- S'il y a répétition de numéros, la valeur est "some-redundancy" ;
- Si aucune des trois situations précédentes n'a été vérifiée, la valeur est "ok".

Les quatre tests sont vérifiés dans cet ordre. Dès que l'un d'eux est positif, la valeur correspondante est assignée à l'attribut *call-sequence* et les tests suivants ne sont pas vérifiés :

```
"définir la valeur de l'attribut call-sequence"

VAR
    ordre_croissant : boolean ;
    missing_notes, repeated_notes : integer ;
    value : string
|[
value := null
; IF missing_notes > 0 →
    value := "incomplete"
[] (missing_notes = 0) AND (ordre_croissant = false) →
```

```
    value := "non-monotonic"
[] (missing_notes = 0) AND (ordre_croissant = true) AND
   (repeated_notes > 0) →
    value := "some-redundancy"
[] (missing_notes = 0) AND (ordre_croissant = true) AND
   (repeated_notes = 0) →
    value := "ok"
FI
]|
```

5.4.2.4 Résumé

Dépendances

La reconnaissance des appels de notes nécessite uniquement le balisage préalable des notes.

Description

Tag-appelnote a pour objectif de baliser les appels de note dans les articles qui contiennent un champ notes. Une première condition vérifie donc la présence de la balise <notes>. L'algorithme recherche ensuite dans l'attribut *count* de <notes> le nombre de notes de l'article (inscrit préalablement par *tag-notes*).

L'algorithme traite d'abord l'entrée et ensuite tous les paragraphes de l'article situés entre l'entrée et le champ <notes>...</notes>. À chaque fois, une chaîne virtuelle est créée, dans laquelle l'algorithme itère sur les balises exposants et y recherche l'expression régulière [1-9][0-9]*\). Lorsqu'il en trouve une, il calcule l'acceptabilité du numéro détecté et regarde si la parenthèse est à l'extérieur ou à l'intérieur de l'exposant. Si la parenthèse est à l'extérieur de l'exposant, les successions de parenthèses sont vérifiées exactement comme expliqué ci-dessus (→ 5.4.2.2). Si la parenthèse est à l'intérieur de l'exposant, ou si l'appel de note candidat passe l'épreuve des successions de parenthèses, il est balisé. Son numéro et son statut d'acceptabilité sont mémorisés.

Dans le cas particulier où deux appels de note candidats ont été trouvés selon le schéma "(...X)...Y)", les deux sont balisés et mémorisés. Dans ce cas, la recherche de l'exposant suivant se fait après le deuxième appel de note (Y) et non le premier (X).

Après avoir traité tous les paragraphes de cette façon, l'algorithme vérifie la correspondance entre les notes et les appels de note. Un attribut *call-sequence* est alors ajouté à la balise <notes> avec pour valeur "ok", "incomplete", "some-redundancy" ou "non-monotonic", et un avertissement est émis.

5.4.3 Balisage de l'entrée : tag-entry

<h>macerare weich machen.</h>	<entry>macerare weich machen.</entry>

(FEW 6/1, 8a, MACERARE)

5.4.3.1 Objectifs

L'algorithme *tag-entry* a pour objectif de baliser le champ de l'entrée. Conformément au modèle proposé (→ 3.5.3), ce champ recouvre

1. dans les matériaux étymologisés, l'étymon-vedette et les informations associées ;
2. dans les matériaux d'origine inconnue (volumes 21-23), le concept sous lequel sont directement regroupés les matériaux.

Outre les raisons mentionnées plus haut (→ 3.5.3), le balisage du champ de l'entrée est nécessaire pour l'application d'autres algorithmes comme *split-doc-com* ou *tag-concept*.

5.4.3.2 Critères de détection

> ITy. Le champ de l'entrée se présente dans une police plus grande que le reste de l'article. Il se termine par un point, sauf exceptions (par exemple FEW 24, 413a, AMBROSINI).
> ITe. Il contient au moins un étymon (volumes 1-20 et 24-25) ou un concept (volumes 21-23).

La détection du champ de l'entrée diffère selon le mode d'acquisition du texte fewien.

En cas de saisie manuelle du FEW, le champ de l'entrée a dû être balisé <h>...</h> dans le document au format FFML (→ 4.3.2). L'algorithme consiste dès lors à repérer ces deux balises, à vérifier qu'elles se situent en début d'article et que la balise </h> est précédée d'un point, enfin à remplacer ces deux balises par les balises <entry>...</entry>.

En cas d'impossibilité de saisie manuelle, la détection du champ de l'entrée est plus complexe : elle dépend fortement de la qualité de la numérisation et, notamment, de la reconnaissance par le logiciel d'océrisation de la taille des polices. Si la taille des polices est bien conservée, il est possible de se servir de cet indicateur pour reconnaître algorithmiquement le champ de l'entrée. Dans le cas contraire, le seul moyen de reconnaissance consiste à détecter la présence d'un étymon ou d'un concept en grasses. Le balisage préalable des étymons (→ 5.4.4) et des concepts (→ 5.4.18) est dès lors requis, ainsi que celui des autres éléments constitutifs du champ de l'entrée (langue d'étymon essentiellement, le balisage de la glose n'étant pas au programme, → 3.5.3). Le début du champ de l'entrée se trouve en tout début du paragraphe où a été détecté l'étymon ou le concept en grasses. La fin du champ de l'entrée se trouve normalement à la fin de ce même paragraphe, juste après un point final.

La méthode de détection du champ de l'entrée en cas de non-saisie manuelle du FEW et de non-fiabilité du logiciel d'océrisation apparaît, en raison des dépendances qu'elle entraîne, comme peu confortable. Tout d'abord, le balisage des étymons et des concepts dépend de listes de mots-clés et n'est donc pas assuré d'exhaustivité

(→ 5.4.4), même s'il est envisageable de mettre en place quelques vérifications pour limiter les erreurs. Ensuite et surtout, les éléments typographiques (mise en grasses, point final) et de mise en page (paragraphes) acquièrent une importance capitale comme indicateurs de reconnaissance, alors que leur reconnaissance par le logiciel d'océrisation est sujette à caution. Le mode d'acquisition du contenu du FEW par saisie manuelle est, ici, clairement préférable.

5.4.3.3 Résumé

Dépendances

La reconnaissance du champ de l'entrée nécessite le balisage préalable de l'article et des paragraphes, ainsi que le balisage des sections de texte apparaissant dans une taille de police plus grande. Ces trois balisages sont prévus dans le format FFML.

En cas de non-saisie manuelle du FEW et de non-fiabilité du logiciel d'océrisation, la reconnaissance du champ de l'entrée dépend du balisage des étymons, des langues d'étymon et des concepts (voir les algorithmes *tag-etymon*, *tag-lang-etymon* et *tag-concept*).

Description de l'algorithme en cas de saisie manuelle du FEW (situation par défaut)

Tag-entry identifie et balise le champ de l'entrée de l'article.

L'algorithme consiste à détecter la présence des balises <h> et </h>, insérées manuellement lors de la saisie du FEW. En cas de détection de moins ou plus d'une balise ouvrante <h> et de moins ou plus d'une balise fermante </h>, le traitement est arrêté. La présence d'un point juste avant la balise fermante </h> est également vérifiée. Après ces vérifications, les balises <entry> et </entry> sont substituées aux balises <h> et </h>.

Description de l'algorithme en cas de non-saisie manuelle du FEW

En cas de non-saisie manuelle du FEW, l'algorithme précédent n'est pas applicable. L'algorithme consiste alors à parcourir le texte à partir du début de l'article en y recherchant les balises <etymon> ou <concept> (selon le numéro de volume). Dès qu'un élément <etymon> ou <concept> est trouvé, sa position au début de l'article est vérifiée, ainsi que la présence proche de balises ... (inclues dans ou incluant l'étymon ou le concept). Une balise ouvrante <entry> est alors insérée au début du paragraphe concerné. La balise fermante </entry> est insérée à la fin de ce même paragraphe, après vérification de la présence d'un point final. Le résultat de cet algorithme est sujet à erreurs, notamment en ce qui concerne la position de la balise fermante </entry>. Un avertissement pour vérification manuelle est obligatoirement émis.

5.4.4 Balisage des étymons : tag-etymon

<entry>hasela (ahd.) art brassen.</entry>	<entry><etymon type="vedette">hasela </etymon> (ahd.) art brassen.</entry>

(FEW 16, 176a, HASELA)

5.4.4.1 Objectifs de l'algorithme

L'algorithme *tag-etymon* a pour objectif de reconnaître et de baliser tous les étymons présents lexicologiquement dans l'article du FEW soumis au logiciel, qu'ils aient lexicographiquement la fonction de lemme (étymon-vedette, étymon de renvoi) ou de sous-lemme (→ 3.5.3 ; 3.7.3). Outre son intérêt direct pour l'exploitation du FEW rétroconverti, le balisage des étymons est nécessaire pour la reconnaissance des renvois, des signatures et des formes (cf. algorithmes concernés).

Conformément au modèle décrit (→ 3.5.3 ; 3.7.3), l'algorithme doit reconnaître la fonction de chaque étymon ("vedette", "renvoi" ou "sous-lemme") et l'indiquer comme valeur d'un attribut *type* associé à la balise <etymon>. Un attribut facultatif *desc* indiquant la descendance de l'étymon ("héréditaire" ou "emprunt") peut être ajouté lorsque l'étymon apparaît dans le commentaire comme reprise de l'étymon vedette :

<etymon type="vedette" desc="héréditaire"><sc>etymon</sc></etymon>

Après application de *tag-etymon*, un article des volumes 1-20 ou 24-25 comporte donc au minimum un élément <etymon type="vedette"> dans l'entrée de l'article. Il peut contenir également d'autres éléments <etymon> dans le reste de l'article (étymons de renvoi, reprise de l'étymon-vedette dans le commentaire, étymons cachés). Les articles des volumes 21-23 n'ont, par définition, pas d'étymon-vedette. Ils peuvent néanmoins comporter des étymons de renvoi ou des étymons cachés (cf. Büchi 1996 11-14), qui doivent être détectés et balisés par *tag-etymon*.

5.4.4.2 Critères de détection

Indicateurs textuels

> ITe. Les étymons vedettes du FEW sont connus. À l'ATILF se trouvent plusieurs listes d'étymons constituées lors de l'élaboration de l'Index (ATILF 2003) et de divers projets tournant autour de l'informatisation du FEW (index onomasiologique et étymons multiples, → 1.5).
> Par ailleurs, Büchi 1996 (annexe 9.1) contient une liste d'étymons cachés (sous-lemmes). Cette liste comporte des étymons présents dans le FEW en tant que lemmes, mais aussi des étymons non mentionnés explicitement, signalés dans la liste par le signe « + ».

Les étymons cachés relevés par Büchi 1996 ont été introduits dans une liste de

mots-clés intitulée *few-etymon-cache-base*[7].

Les différentes listes d'étymons-vedettes existantes ont été fusionnées en une seule, intitulée *few-etymon-base*. La fusion a montré d'énormes divergences au niveau du contenu : dans le nombre d'étymons répertoriés, dans le traitement de l'astérisque (parfois conservé, parfois pas) et dans le traitement des étymons multiples (cf. Büchi 1996, 80), tantôt conservés tels quels (par exemple *ulwo-, -a* [FEW 14, 16a]), tantôt explicités (*ulwo* et *ulwa*), tantôt simplifiés (*ulwo* uniquement). Des disparités ont également été constatées en ce qui concerne le codage des caractères, non normalisé d'une liste à l'autre. Un travail de nettoyage a donc été entrepris. La liste finale comporte environ 20 400 étymons codés en UTF-8 selon la table des caractères du FEW (→ 4.2.2).

Il est important de noter que la liste *few-etymon-base* n'est pas une copie conforme de toutes les formes d'étymon-vedette tels qu'elles apparaissent dans le FEW. Les étymons pourvus d'une flexion (étymons complexes comportant une désinence, par exemple CHILIAS, -ADIS [FEW 2, 636a], et étymons multiples tels que ULWO-, -A) sont répertoriés comme des formes simples. En effet, l'inconvénient d'une recherche par mot-clé est qu'elle ne permet aucune variante : un mot-clé tel que *ulwo-, -a* ne permet de détecter ni *ulwo-,-a* (sans espace), ni *ulwo, -a* (sans tiret), ni même *ulwo*. En revanche, le mot-clé *ulwo* permet de détecter chacune de ces trois séquences. La répertorisation sous les formes simples permet donc d'éviter les faux négatifs.

Malgré cette précaution, les indicateurs textuels ne suffisent pas pour reconnaître les étymons du FEW. Ils génèrent en effet beaucoup de faux positifs. Des items comme *Marivaux* ou *bord*, qui appartiennent aux mots-clés de *few-etymon-base*, n'ont pas nécessairement le statut d'étymons chaque fois qu'ils apparaissent dans le texte fewien. Il est donc nécessaire de compléter les indicateurs textuels par d'autres types d'indicateurs.

[7]La numérisation des pages concernées dans Büchi 1996 et l'établissement de la liste dans le format requis ont été réalisées à l'ATILF par Isabelle Clément, que nous remercions.

Indicateurs typographiques et positionnels

> IPo/ITy. Les indicateurs typographiques et positionnels diffèrent selon la fonction de l'étymon dans le FEW.
>
> – L'étymon-vedette de l'article se trouve en grasses dans le champ de l'entrée. En dehors de l'entrée, il se présente soit en petites capitales, si sa descendance est (au moins en partie) héréditaire, soit en italiques, si sa descendance est constituée uniquement d'emprunts.
>
> – Un étymon de renvoi se trouve toujours en petites capitales, sauf dans le champ de l'entrée, où il peut apparaître en grasses (cf FEW 6/1, 1b, MACCARE ; 16, 296b, KAJUIT ; 18, 6b, ALLIGATOR).
>
> – Les étymons cachés (sous-lemmes), lorsqu'ils sont explicités, se présentent le plus souvent en italique, mais aussi en petites capitales (cf. ALBA dans FEW 1, 63a, ALBUS).
>
> Ces règles typographiques et positionnelles valent pour tous les articles, excepté ceux rédigés par Zumthor (cf. Büchi 1996, 139). Dans ces derniers, la reprise de l'étymon-vedette dans le commentaire s'effectue en petites capitales même si sa descendance est constituée d'emprunts uniquement (cf. FEW 4, 102b, GENITIVUS ; 755b, INTERJECTIO).

Les indicateurs typographiques permettent :

1. de distinguer les vrais étymons des faux (les faux étymons étant des mots-clés qui ne sont pas en grasses, en italique ou en petites capitales) ;

2. associés avec les indicateurs positionnels, de définir le type de l'étymon (vedette, renvoi ou sous-lemme) ;

3. dans certains cas (étymon-vedette positionné hors entrée, dans un article non rédigé par P. Zumthor), de déterminer si la descendance de l'étymon est héréditaire ou empruntée.

Combinaison des indicateurs

À partir des deux listes d'étymons constituées et des indicateurs typographiques, il est possible d'élaborer un algorithme de détection des étymons. La règle de base est la suivante : un extrait du FEW est un étymon s'il correspond à un mot-clé des listes *few-etymon-base* ou *few-etymon-cache-base* et qu'il est entouré des balises , <i> </i> ou <sc> </sc>.

Cas particuliers

> ITy. Dans les étymons complexes tels que « ulwo, -a », les suffixes flexionnels ou dérivationnels ne sont pas nécessairement en grasses, mais aussi en italiques ou en romaines : les pratiques varient.
> Les étymons reconstruits sont précédés d'un astérique qui n'est pas nécessairement en grasses ou ne sera pas balisé comme tel lors de la saisie.

La combinaison des indicateurs textuels et typographiques pose un problème important : la non-concordance éventuelle entre le mot-clé et le balisage de formatage. Le mot-clé peut dépasser le balisage, par exemple en cas d'étymons avec astérisque, tel que "*bla", qui peuvent être transcrits dans le document FFML comme "*bla", mais aussi "*bla". De même, un mot-clé peut ne constituer qu'une partie du texte en grasses dans les cas de formes flexionnelles : etymon, -désinence.

Le problème est résolu de deux manières :

- par l'algorithme de prétraitement streamline-layout-tags, qui rétablit les astérisques à l'intérieur des balises de formatage ;
- par le recours, dans *tag-etymon*, à une méthode qui détecte les balises typographiques dans l'entourage du mot-clé et non juste à côté de celui-ci.

5.4.4.3 Méthode de détection des balises typographiques

Afin de résoudre le problème de non-imbrication des balises avec le mot-clé, la recherche des balises attenantes au mot-clé est transformée en une recherche des balises <sc>, et <i> les plus proches autour du mot-clé. Cela signifie qu'il n'est pas nécessaire que la balise typographique soit collée au mot-clé. Concrètement, l'algorithme cherche la balise ouvrante <sc>, ou <i> la plus proche à gauche du mot-clé, puis la balise fermante </sc>, ou </i> la plus proche à droite du mot-clé. Il compare ensuite les deux balises trouvées :

- s'il n'en a trouvé aucune : il conclut qu'il n'est pas en présence d'un étymon ;
- s'il n'en a trouvé qu'une ou qu'elles ne correspondent pas : il émet une alerte ;
- si elles correspondent, il conclut qu'il est en présence d'un étymon.

5.4.4.4 Traitement de l'entrée

> IPo. Dans le champ de l'entrée peuvent se trouver un ou plusieurs étymons-vedettes (plusieurs en cas d'étymons multiples) et un ou plusieurs étymons de renvoi.
> ITy. Les étymons-vedettes se présentent en grasses ; les étymons de renvoi sont en grasses ou en petites capitales.

Ces étymons sont censés se trouver dans la liste *few-etymon-base*. L'algorithme consiste donc à rechercher dans l'entrée les mots-clés de *few-etymon-base* et à les fil-

trer pour conserver uniquement ceux qui se trouvent entre les balises ... ou <sc>...</sc>.

Si après filtrage, il n'en reste aucun ou s'il en reste plus de trois, une alerte est émise. Si un, deux ou trois étymons ont été trouvés, l'algorithme détermine leur type selon la balise typographique trouvée : la valeur est « vedette » si la balise est <b> et « renvoi » si la balise est <sc>. Les étymons de renvoi en grasses sont donc typés provisoirement comme des vedettes, ce qui ne pose pas de problème à cette étape de la rétroconversion : les étymons de renvoi seront en effet reconnus plus tard par *tag-renvoi*. Le résultat est soumis à une analyse de conformité, de façon à respecter la règle qui veut que le premier étymon soit une vedette, que le deuxième étymon éventuel soit une vedette ou un renvoi, et que le troisième étymon éventuel soit un renvoi. L'algorithme n'accepte donc que les scénarios suivants :

– une vedette seule ;
– deux vedettes (étymons multiples) ;
– une vedette et un renvoi (en grasses ou en petites capitales) ;
– deux vedettes (étymons multiples) et un renvoi en petites capitales.

5.4.4.5 Traitement des paragraphes

> IPo. Dans le reste de l'article, et notamment dans le commentaire et les notes, peuvent apparaître des étymons qui remplissent différentes fonctions : soit celle de l'étymon-vedette de l'article, repris dans le commentaire ; soit celle d'un étymon de renvoi ; soit celle d'un étymon caché (sous-lemme).
> ITy. Hors entrée, les étymons de renvoi sont en petites capitales uniquement. Les étymons cachés et les reprises de l'étymon-vedette sont en petites capitales ou en italiques.

Les reprises de l'étymon-vedette et les étymons de renvoi trouvent leur correspondant dans *few-etymon-base*. Les étymons cachés sont répertoriés dans *few-etymon-cache-base*. Les étymons qui n'apparaîtraient dans aucune base seront repérables manuellement via l'algorithme de post-traitement show-untagged-text (→ 5.5.1).

Le traitement préalable de l'entrée a permis de connaître l'étymon-vedette. Les reprises de ce dernier dans le commentaire seront donc repérables par simple comparaison de mots.

L'algorithme cherche d'abord dans le paragraphe tous les mots-clés de *few-etymon-base*. Il compare chaque mot-clé trouvé à l'étymon vedette de l'article (préalablement balisé lors du traitement de l'entrée). Si le mot-clé est une reprise de la vedette, il le balise avec un attribut *type*="vedette" et un attribut *desc*, qui reçoit la valeur "héréditaire" ou "emprunt" en fonction de la balise typographique. Si ce n'est pas la vedette, il vérifie que l'étymon est bien en petites capitales et il balise le mot-clé avec un attribut *type*="renvoi". Si l'étymon est en italiques, il est balisé avec un attribut *type*=" ?".

L'algorithme cherche ensuite les mots-clés de *few-etymon-cache-base*. Chaque mot-clé trouvé est balisé avec un attribut *type* recevant la valeur "sous-lemme".

5.4.4.6 Résumé

Dépendances

Tag-etymon utilise deux listes de mots-clés : *few-etymon-base* et *few-etymon-cache-base*. Il dépend des algorithmes *streamline-layout-tags*, *detect-dubious-spacing* et *tag-entry*.

Remarque. En cas d'impossibilité de saisie manuelle du FEW, *tag-etymon* ne peut profiter du balisage préalable de l'entrée. Cette situation problématique entraîne des modifications profondes de l'algorithme et de sa place dans la séquence de rétroconversion (→ 5.4.3).

Description

Tag-etymon identifie et balise dans un article du FEW les étymons répertoriés dans les listes *few-etymon-base* et *few-etymon-cache-base*. Il ne traite pas l'entrée des articles des volumes 21 à 23.

Chaque étymon balisé reçoit obligatoirement un attribut *type*, qui peut recevoir les valeurs "renvoi", "sous-lemme", "vedette" ou " ?". Les étymons qui reprennent la vedette reçoivent également un attribut *desc*, qui peut recevoir les valeurs "héréditaire" ou "emprunt". La valeur de chaque attribut est déterminée après détection des balises typographiques incluant l'étymon (voir ci-dessus).

Le traitement de l'entrée s'effectue dans une chaîne virtuelle où la plupart des balises sont transparentes (cf. partition). Pour chaque mot-clé de *few-etymon-base* détecté, l'attribut *type* est d'abord calculé. Les mots-clés qui n'ont pas pu recevoir de type (à cause de l'absence de balises ou <sc>) ne sont pas balisés, les autres sont balisés <etymon> et reçoivent leurs attributs. S'il y a moins d'un étymon ou plus de trois étymons, une alerte est émise et le traitement est arrêté pour vérification manuelle. Le type des étymons est ensuite vérifié : le premier doit avoir « vedette » pour valeur ; le type du deuxième étymon doit être "vedette" ou "renvoi" ; celui du troisième doit être "renvoi". Si ce n'est pas le cas, une alerte est émise.

Le traitement d'un paragraphe s'effectue en deux phases : tout d'abord la détection et le balisage des mots-clés de *few-etymon-cache-base*, ensuite la détection et le balisage des mots-clés de *few-etymon-base*. Chaque phase s'effectue après création d'une chaîne virtuelle (cf. partition) dans laquelle les balises <etymon> sont rendues invisibles, de sorte que les étymons balisés lors de la première phase ne sont pas traités lors de la seconde.

Lors de la première phase, les mots-clés de *few-etymon-cache-base* identifiés reçoivent « sous-lemme » comme valeur de l'attribut *type*.

Lors de la seconde phase, les mots-clés de *few-etymon-base* identifiés dans le paragraphe sont comparés à l'étymon-vedette de l'article (balisé lors du traitement de l'entrée), de façon à définir leur attribut *type* et (uniquement si *type* reçoit "vedette" pour valeur) leur attribut *desc* (cf. ci-dessus).

5.4.5 Balisage des langues d'étymon : tag-lang-etymon

<entry><etymon type="vedette"> hasela </etymon> (ahd.) art brassen.</entry>	<entry><etymon lang="ahd." type="vedette"> hasela</etymon> <lang_etymon>(ahd.) </lang_etymon> art brassen.</entry>

(FEW 16, 176a, HASELA)

5.4.5.1 Objectifs de l'algorithme

L'algorithme *tag-lang-etymon* a pour objectif de reconnaître la langue à laquelle appartient l'étymon-vedette de l'article, de façon à permettre la recherche de lexèmes selon leur origine linguistique lors de la phase d'exploitation du FEW.

Lorsque la langue est mentionnée explicitement entre parenthèses à la droite de l'étymon-vedette, *tag-lang-etymon* doit la baliser <lang_etymon>[8], en incluant les parenthèses, conformément à la modélisation proposée ($\rightarrow$ 3.5.3). *Tag-lang-etymon* doit également ajouter le nom de cette langue comme valeur d'un attribut *lang* assigné à l'élément <etymon> qui identifie l'etymon-vedette (préalablement balisé par *tag-etymon*).

Si aucune mention de langue n'apparaît dans l'entrée de l'article, *tag-lang-etymon* n'effectue aucun balisage, mais il doit malgré tout déterminer la langue implicite de l'étymon-vedette et l'attribuer à l'élément <etymon>, de la même façon que ci-dessus. Les onomatopées et les éponymes – qui, bien qu'appartenant non moins que le reste du lexique à un système linguistique particulier, ne bénéficient pas de marquage linguistique dans le FEW – doivent être identifiés comme tels.

Tag-lang-etymon ne balise pas les indications de langue d'étymons qui apparaissent dans le commentaire ou les notes du FEW. Ce balisage est en effet assuré par un autre algorithme (cf. *tag-geoling*).

5.4.5.2 Critères de détection

Indicateurs textuels

Est-il possible d'établir un lexique exhaustif de toutes les langues d'étymon qui apparaissent dans le FEW ?

> ITe. Büchi 1996, 26-38 énumère les diverses origines linguistiques des étymons du FEW, mais sans citer systématiquement les abréviations exactes telles qu'elles apparaissent dans le FEW. Matthey et Thibault en préparation propose une liste de ces abréviations, mais qui comprend quelques oublis.

Les listes existantes ne sont pas exhaustives et, surtout, sont susceptibles de varia-

[8] L'élément <lang_etymon> ne contient aucun attribut.

tion par rapport à ce qui se trouve exactement dans le FEW. Nous avons donc préféré parcourir le FEW pour établir une nouvelle liste, exhaustive, des langues qui accompagnent les étymons-vedettes. La liste résultante, intitulée *few-lang-etymon-base*, comprend toutes les langues telles qu'elles apparaissent dans le champ de l'entrée (c'est-à-dire sous leur forme abrégée et entre parenthèses).

> La langue de l'étymon est parfois multiple (FEW 19, 16a, BADAWARD : [pers. ar.]).

Les mentions de langues multiples sont répertoriées telles quelles dans *few-lang-etymon-base*. Par conséquent, elles ne posent aucun problème algorithmique.

Indicateurs positionnels

> IPo. La langue de l'étymon-vedette fait partie du champ de l'entrée.
> Lorsqu'elle apparaît explicitement, elle se trouve généralement juste après l'étymon-vedette (des exceptions, rares, ont été relevées par exemple dans le volume 25, ainsi 619a : cf. Büchi 1996, 80, n. 7).
> ITy. Elle apparaît toujours entre parenthèses.

Les indicateurs positionnels et typographiques ne permettent pas de reconnaître avec spécificité qu'un item représente une langue d'étymon. La règle selon laquelle « toute chaîne de caractères qui se trouve dans le champ de l'entrée entre parenthèses est une langue d'étymon » génère en effet de nombreux faux positifs : entre parenthèses dans l'entrée apparaissent aussi des catégories grammaticales (FEW 3, 646b, FLUXUS), des éléments de glose (FEW 2/1, 489b, CATALOGUS) ou des parties d'étymon (FEW 2, 31a, CADIÈRE ; 9, 462b, PRO RATA).

Combinaison des indicateurs

Les indicateurs textuels apparaissent comme fiables. Ils sont en outre assez spécifiques pour distinguer avec certitude les mentions explicites de langues d'étymon-vedette, à condition que les mots-clés intègrent les parenthèses englobantes et qu'ils soient recherchés dans le champ de l'entrée uniquement (indicateur positionnel).

Il est à remarquer que la reconnaissance des langues d'étymon ne nécessite pas la reconnaissance préalable des étymons.

5.4.5.3 Traitement de l'implicite

> Lorsqu'aucune mention de langue n'apparaît dans le champ de l'entrée, la langue de l'étymon-vedette est implicitement celle de la section linguistique dans laquelle se trouve l'article (cf. Büchi 1996, 80-83). En pratique, seules les sections 1, 3 et 5 sont concernées par cet implicite. Dans la section 1 (volumes 1-14 et 24-25), la langue implicite est le latin. La section 3 (volume 18) comporte uniquement des étymons-vedette en anglais moderne. Enfin, la section 5 (volume 20) contient plusieurs sous-sections, dont certaines seulement sont concernées par l'implicite. Les langues implicites sont : le breton (20, 1-17 ; 116), le basque (20, 18), l'hébreu (20, 24-28), le tsigane (20, 29-30), le hongrois (20, 31-32), l'eskimo (20, 54) et l'australien (20, 114).

Dans ces sections, les numéros de volume et de page correspondant à l'article permettent donc de déduire la langue implicite de l'étymon-vedette.

> Les seules incohérences relevées par Büchi 1996, 82 concernent les étymons-vedettes DOHNE (3, 115b), DAWAMESK (3, 20b) et DANISK (15/2, 53b), qui, de façon erronée par rapport à la grammaire du FEW, ne présentent aucune mention explicite de langue.

Ces rares erreurs seront à corriger manuellement après rétroconversion.

5.4.5.4 Cas particuliers : onomatopées et noms propres

Deux catégories d'étymons-vedette n'ont pas de "langue d'appartenance" (cf. Büchi 1996, 81), mais contiennent malgré tout une indication sur leur catégorie.

> ITe. Les étymons onomatopéiques sont rarement signalés explicitement par une indication telle que *lallwort* ou *schallwort* (FEW 8, 571). Les étymons onomastiques sont parfois (mais pas toujours) signalés par une indication telle que *personenname* (FEW 1, 329), *NP* (« nom de personne »), *ortsname*, *NL* (« nom de lieu ») ou *völkername* ; on trouve aussi des mentions telles que *stadt* (FEW 1, 333, BERLIN) ou *hauptstadt* (FEW 18, 80, LONDRES).
> Büchi 1996 contient un index des étymons onomatopéiques (392-393) et un index des éponymes (564-589).
> IPo. Ces indications explicites apparaissent entre parenthèses ou non.

L'absence de règles de présentation systématiques rend peu intéressant le balisage des mentions explicites. En revanche, il est utile de signaler les étymons onomatopéiques ou onomastiques pour l'exploitation future du FEW. En ce qui concerne les étymons onomatopéiques, l'index élaboré par Büchi 1996 constitue un indicateur suffisant. Une liste *few-onomatop-base* a été créée. Lorsque l'étymon-vedette appartient à cette liste, il reçoit un attribut *lang*="onomatopee".

Les éponymes sont quant à eux reconnaissables par le fait qu'ils commencent par une majuscule. Un étymon-vedette sera donc implicitement un nom propre, quelle que

soit la section linguistique dans laquelle il apparaît, s'il commence par une majuscule (astérisque exclu). Tout étymon commençant par une majuscule recevra un attribut *lang*="nom_propre".

L'utilisation de l'attribut *lang* pour signaler les éponymes et les étymons onomatopéiques se justifie uniquement parce que le discours lexicographique traite ces deux informations, logiquement indépendantes, sur le même pied que l'appartenance à un système linguistique.

Il est à remarquer que la reconnaissance des étymons onomatopéiques et onomastiques nécessite le balisage préalable des étymons-vedette.

5.4.5.5 Résumé

Dépendances

Tag-lang-etymon dépend de l'algorithme de prétraitement *detect-dubious-spacing* ainsi que du balisage de l'étymon-vedette (*tag-etymon*).

Il utilise deux bases de mots-clés : *few-lang-etymon-base* et *few-onomatop-base*.

Description

L'algorithme *tag-lang-etymon* détecte et balise la langue de l'étymon-vedette d'un article du FEW. Il détermine également si l'étymon est une onomatopée ou un nom propre.

La détection a lieu dans le champ de l'entrée uniquement (cf. *tag-entry*). Le traitement a lieu sur une chaîne virtuelle ayant </entry> comme balise terminale. L'algorithme recherche dans la chaîne virtuelle les mots-clés qui se trouvent dans *few-lang-etymon-base*, et le mot-clé trouvé est balisé. Si plus d'un mot-clé est trouvé, l'algorithme émet une alerte et arrête le traitement.

Un attribut *lang* est ajouté à l'étymon-vedette. La valeur de cet attribut est donnée par la langue d'étymon explicite qui a été reconnue et balisée. Si aucune indication de langue explicite n'a été reconnue, l'algorithme cherche si l'étymon commence par une majuscule (auquel cas l'attribut *lang* reçoit la valeur "nom_propre") ou s'il appartient à la liste *few-onomatop-base* (auquel cas il reçoit la valeur "onomatopee"). Si aucun de ces deux tests n'est positif, l'algorithme attribue à l'étymon la langue implicite donnée par les numéros de volume et de page de l'article.

5.4.6 Balisage des renvois internes : tag-renvoi

<art book="1" ici="2" id="20" volume="3"> S. <etymon type="renvoi"> <sc>trinken</sc></etymon>.	<art book="1" ici="2" id="20" type="renvoi" volume="3"> <renvoi>S. <etymon type="renvoi"> <sc>trinken</sc></etymon></renvoi>.

(FEW 3, 160a, DRINKEN)

5.4.6.1 Objectifs

Tag-renvoi a pour objectif de reconnaître et de baliser les renvois internes au FEW, mais externes à l'article, c'est-à-dire les renvois vers un autre article du FEW[9]. Ces renvois sont balisés, conformément à la modélisation proposée (→ 3.5.4), par l'élément <renvoi>.

Tag-renvoi doit également détecter si l'article est un article de renvoi (Büchi 1996, 131-133 ; → 3.5.4) et, le cas échéant, ajouter à la balise <art> un attribut *type* recevant la valeur "renvoi". Ce traitement doit permettre aux algorithmes suivants de distinguer les articles de renvoi des articles avec contenu.

Tag-renvoi ne traite pas les renvois vers les articles des volumes 21 à 23 : ces renvois à des concepts sont balisés par *tag-concept* (→ 5.4.18).

5.4.6.2 Reconnaissance des articles de renvoi

> IPo. Les articles de renvoi sont reconnaissables par le fait qu'ils ne contiennent pas de documentation, mais uniquement un renvoi (FEW 3, 160a, DRINKEN).
> ITe. Un relevé des étymons concernés a été élaboré à l'ATILF par Gilles Souvay. Ce relevé n'est toutefois pas exhaustif[10].

Nous avons normalisé le relevé existant pour qu'il corresponde au codage requis. La liste résultante, nommée *few-renvoi-base*, contient 217 étymons-vedettes d'articles de renvoi.

> Les étymons multiples posent problème. FEW 18, 72b, HUMOR renvoie par exemple à FEW 4, 513, HUMOR. De ces deux étymons identiques, seul le premier (FEW 18, 72b) est la vedette d'un article de renvoi.

Les indicateurs textuels sont suffisants pour détecter les articles de renvoi, à condition de prendre en compte les étymons multiples dont l'un seulement est un renvoi. L'algorithme consiste à chercher dans l'entrée de l'article les étymons (préalablement balisés par *tag-etymon*) et à regarder si le premier d'entre eux est un mot-clé de *few-renvoi-base*. Le cas échéant, quelques vérifications sont effectuées afin de s'assurer qu'il s'agit bien d'un article de renvoi, ensuite le renvoi est balisé et l'article est typé comme article de renvoi.

Les vérifications portent sur le nombre de paragraphes et sur la présence d'éléments <etymon>. Si l'article contient plus d'un paragraphe (entrée exclue), l'article n'est pas considéré comme un article de renvoi. Si l'article contient un seul paragraphe en dehors de l'entrée, il est considéré comme un article de renvoi uniquement si le paragraphe en question commence par « S. » et contient un étymon. Si l'article ne contient aucun paragraphe, l'article est considéré comme un article de renvoi, le renvoi étant constitué par le deuxième étymon trouvé dans l'entrée ; si l'entrée contient un seul étymon,

[9]Pour les renvois externes au FEW, voir le balisage des références bibliographiques (→ 5.4.9). Pour les renvois internes à l'article, voir le balisage du marquage alphanumérique (→ 5.4.12).

[10]Il y manque notamment l'étymon HADDOCK (FEW 18, 70b), qui renvoie à FEW 16, 110, HADDOCK.

il s'agit d'une situation anormale qui justifie un arrêt du traitement pour vérification manuelle.

5.4.6.3 Reconnaissance des renvois hors articles de renvoi

> IPo. Les renvois peuvent apparaître à tout endroit de l'article. Ils se présentent sous des formes très variées. En voici quelques exemples :
>
> « (s. hier 6, I, 122) » (FEW 24, 2a)
> « Hier 1, 4 » (FEW 24, 26b)
> « [...] * BUR, s. hier bd. 15, II » (FEW 24, 26b)
> « s.v. NUCICULA hier 7, 226 » (FEW 25, 10b)
> « ici 1, 341b, BESTIA II 1 » (FEW 25, 961a)
> « Cf. infra AUSTERUS 4b » (FEW 25, 1070)
> « S. noch AMATOR. » (FEW 18, 7a)
> « vgl. hier 1, 152 *ASCIATA » (FEW 20, 5b)

La difficulté principale, dans l'élaboration de l'algorithme de reconnaissance des renvois, consiste à trouver un motif qui corresponde à la grande variété de formes que présente le FEW. La définition même du renvoi interne (« indication servant à renvoyer le lecteur vers une autre partie du FEW ») permet de définir un renvoi comme une expression comportant deux sortes d'éléments :

- d'une part, des marqueurs spécifiques indiquant qu'il s'agit d'un renvoi ;
- d'autre part, des informations sur la cible vers laquelle dirige le renvoi.

L'analyse des exemples confirme cette définition : la combinaison d'un marqueur et d'une cible indique un renvoi, de façon à la fois fiable et spécifique. Il reste à définir précisément quels sont les marqueurs et les cibles, ainsi que leurs différentes possibilités de combinaison.

Définition des marqueurs et des cibles

> Les marqueurs utilisés par le FEW pour signaler un renvoi sont les suivants :
>
> 1. la lettre s suivie d'un point (abréviation de l'allemand *siehe*) ou, dans la refonte, *s.v.* (abréviation de *sub verbum*) ;
> 2. l'adverbe allemand *hier* ou sa traduction française *ici* (signifiant « dans le FEW ») ;
> 3. les adverbes latins *supra* ou *infra*, utilisés uniquement dans les volumes 24 et 25 de la refonte.
>
> Pour citer la cible, le FEW dispose de deux moyens :
>
> 1. la mention de l'étymon-vedette de l'article cible, éventuellement suivie du marquage alphanumérique indiquant la partie de l'article particulièrement concernée ;
> 2. une « référence FEW », c'est-à-dire le volume, la page et éventuellement la colonne où se trouve la cible.

Définition des combinaisons

À partir de ces cinq éléments, trois concernant le marqueur et deux la cible, et sachant que la présence au minimum d'un marqueur et d'une cible est nécessaire, il est possible de définir six combinaisons minimales permettant de reconnaître un renvoi de façon à la fois fiable et spécifique :

1. *s.* + référence FEW
2. *hier*/*ici* + référence FEW
3. *supra*/*infra* + page
4. *s.*/*s.v.* + étymon (ou l'inverse : étymon + *s.*/*s.v.*)
5. *hier*/*ici* + étymon (ou l'inverse : étymon + *hier*/*ici*)
6. *supra*/*infra* + étymon

Les combinaisons 3 et 6 sont particulières : elles apparaissent dans la refonte uniquement ; en outre, puisque ce marqueur est réservé à une référence dans le même volume, la référence FEW est limitée à une indication de page.

À chacune de ces combinaisons minimales peuvent s'ajouter un ou plusieurs des autres constituants, ce qui en pratique permet un nombre important de combinaisons. Ces combinaisons maximales doivent être prises en compte dans l'algorithme si l'on veut que le balisage encadre bien toute l'expression de renvoi (et non seulement les deux constituants reconnus). Il est intéressant par exemple d'intégrer dans le balisage

du renvoi le marquage alphanumérique qui suit éventuellement l'étymon, afin de le distinguer du marquage interne à l'article (cf. *tag-numbering*, → 5.4.12).

Cas particulier : les renvois multiples

Les renvois multiples présentent un cas particulier de renvois. Suivant la règle d'économie voulant qu'une information déjà citée ne soit pas répétée, ces successions de renvois présentent la mention de plusieurs cibles (étymons et références FEW) avec non-répétition systématique des marqueurs. Dans l'exemple ci-dessous, les nombres entre parenthèses après les étymons sont interprétables comme des renvois vers un autre endroit du FEW uniquement parce que le premier étymon est suivi du marqueur de renvoi explicite *siehe* :

> « S. noch CANIS (hier 2, 194), COHORS (2, 850), COR (2, 1176), CORPUS (2, 1214) [. . .] » (FEW 24, 5a)

En outre, l'ordre de succession des informations concernant les cibles n'est pas systématique. Dans l'exemple ci-dessous, l'étymon et la référence FEW correspondante sont séparés :

> « S. noch JOCUS, GRÎS, hier 4, 44a ; 16, 81. » (FEW 20, 12b)

5.4.6.4 Traduction des constituants au moyen d'expressions régulières

Mettons à part pour l'instant les renvois multiples, et analysons plus en détail les expressions de renvoi telles qu'elles peuvent se présenter dans la version informatique du FEW. Notre objectif final est de définir des expressions régulières qui (1) prennent en compte toutes les variantes et subtilités typographiques du FEW imprimé et (2) englobent tous les constituants présents dans une même expression de renvoi (combinaisons maximales). Pour définir des expressions régulières maximales, il est nécessaire de définir d'abord correctement les constituants.

Le marqueur « s. »

« s. » peut apparaître en majuscule (*S.*) en début de phrase : l'expression régulière adéquate est donc `/[sS]\./`. Nous représenterons dorénavant cette expression régulière au moyen d'une variable dénommée $s.

Le marqueur « s.v. » est quant à lui représenté par l'expression régulière `/[sS]\.[vV]\./`, dénommée $sv.

Le marqueur « hier » / « ici »

L'expression régulière représentant toutes les variantes du marqueur « hier » / « ici » est `/([hH]ier|[iI]ci)/`, symbolisée par la variable $hier.

Le marqueur « supra » / « infra »

L'expression régulière représentant toutes les variantes du marqueur « supra » / « infra » est `/([sS]upra|[iI]nfra)/`. Nous représenterons dorénavant cette expression

régulière par la variable $supra.

L'étymon-cible

L'étymon-cible présent dans une expression de renvoi n'est pas exprimé au moyen d'une expression régulière. Les étymons sont en effet représentés au moyen d'une liste de mots-clés (*few-etymon-base*, cf. *tag-etymon*). L'application de *tag-etymon* avant *tag-renvoi* permet en outre de les reconnaître directement grâce à leur balisage.

La référence FEW

Ce constituant est complexe. Il se divise essentiellement en deux constituants, la mention de volume et la mention de page(s), avec des possibilités de variantes.

Le volume est constitué d'un nombre entre 1 et 25, puis éventuellement d'une indication de tome en chiffres romains (de I à III). Entre le volume et le tome peuvent se trouver une barre oblique, une virgule ou/et un espace, ce qui donne l'expression régulière suivante :

```
/([1-9]|1[0-9]|2[0-5])([\/,_]+(I|II|III))?/
```

Nous schématisons cette expression régulière en une variable $vol.

Cette mention de volume peut être précédée, mais rarement, de l'indication "bd." ou (en français) "t.". Comme cette indication ne peut apparaître dans toutes les combinaisons de renvois, nous préférons ne pas l'intégrer dans la même expression régulière que $vol. Nous créons une nouvelle expression, représentée par la variable $tome :

```
/((bd)|(t))\._?$vol/
```

La page est constituée d'un nombre de 1 à 4 chiffres. Il peut s'agir aussi d'un groupe de pages, un tiret séparant la première de la dernière (567-571). L'expression régulière est donc la suivante :

```
/[0-9]{1,4}(_?-_?[0-9]{1,4})?/
```

Il faut y ajouter la colonne, constituée d'un "a" ou d'un "b", ce qui s'exprime `[ab]`. La mention de la colonne suit celle de la page. Dans le cas d'un groupe de pages, il faudrait donc insérer la colonne à deux endroits de l'expression régulière, et la rendre facultative, comme ceci :

```
/[0-9]{1,4}_?[ab]?(_?-_?[0-9]{1,4}_?[ab]?)?/
```

Nous retenons cette expression et nous la représentons par la variable $page. Nous créons en outre une variable $pagecol, similaire à $page mais qui rend obligatoire la mention de la colonne :

```
/[0-9]{1,4}_?[ab]_?(-_?[0-9]{1,4}_?[ab])?/
```

Les constituants ci-dessus ayant été définis et représentés symboliquement, nous pouvons définir une référence FEW. Celle-ci se compose, soit de $tome seul (comprenant $vol) sans $page, soit de $vol suivi de $page. Ce qui donne l'expression régulière suivante :

```
/(($tome)|($vol)[,_]+($page))/
```

Nous représentons cette expression régulière par la variable $reffew.

5.4.6.5 Traduction des combinaisons au moyen d'expressions régulières

Maintenant que nous avons défini les constituants d'un renvoi FEW, nous pouvons tenter d'exprimer les combinaisons maximales admises, ce qui doit permettre de reconnaître et de baliser les expressions de renvoi dans leur totalité. En effet, entre le marqueur et la cible peuvent se trouver plusieurs autres éléments.

Parmi les six combinaisons définies plus haut, celles qui activent le constituant "étymon" sont particulières, puisque l'étymon ne peut être intégré dans une expression régulière. Ces combinaisons sont exprimées au moyen de deux expressions régulières, distinguant ce qui précède l'étymon et ce qui le suit.

hier/ici + référence FEW

Entre le marqueur ($hier) et la référence ($reffew) peuvent se trouver une virgule et un espace. L'expression régulière est donc `/$hier[,_]*$reffew/`.

En combinaison maximale, cette expression peut recevoir comme autres constituants le marqueur s./S., éventuellement suivi d'un mot tel que *noch*, le tout placé avant le marqueur *hier* ou *ici* :

> « S. noch hier 18, 1 » (FEW 24, 38a)

L'expression régulière maximale intègre donc le constituant $s, suivi éventuellement d'un mot, ce qui peut être représenté de la façon suivante :

```
/($s(_[a-zA-Z]*)?_?)?$hier[,_]?$reffew/
```

Nous désignerons dorénavant cette expression par la variable $regexRenvoiSansEtymon1.

s. + référence FEW

Comme la précédente, cette expression peut recevoir un mot tel que *noch*, placé après le marqueur *s./S.*. Par ailleurs, la référence FEW ne peut contenir d'indication de tome, mais uniquement une indication de volume suivie d'une indication de page. Le tout s'exprime ainsi :

```
/$s(_$mot)?_?$vol[,_]+$page/
```

Cette expression est désignée par la variable $regexRenvoiSansEtymon2.

Le marqueur $s pose problème, car il est utilisé non seulement pour les renvois, mais aussi pour les références bibliographiques. Il peut donc y avoir ambiguïté lorsque les numéros de volume et page ressemblent à celles du FEW, comme dans l'exemple ci-dessous, qui correspond à notre combinaison de renvoi *s.* + référence FEW (rappelons qu'entre le $s et $reffew, nous avons permis la présence d'un mot, afin de pouvoir baliser des expressions comme « s. noch 3, 117 ») :

> « s. Schwyzer 1, 426 » (FEW 18, 3a)

Bien sûr, le sigle de l'ouvrage indique de suite au lecteur qu'il ne s'agit pas du FEW. Pour lever l'ambiguïté dans un traitement informatique, il faudrait donc que les sigles bibliographiques aient été balisés avant les renvois. Même dans ce cas, il est possible que certains sigles ne soient pas reconnus. Une autre solution serait de restreindre l'ex-

pression régulière définie plus haut de façon à ce qu'elle ne puisse s'appliquer qu'au FEW. Cette deuxième option peut se faire, soit en définissant exactement les mots qui peuvent s'intercaler entre $s et $reffew (comme *noch* ou *parallelen* pour citer les plus fréquents) – ce qui est impossible sans lire tout le FEW – soit en obligeant la mention de la colonne – ce qui semble effectivement la règle, mais empêche de reconnaître des exceptions. Pour plus de couverture, nous pouvons combiner les deux options et accepter comme renvoi FEW les expressions qui répondent à l'une des deux définitions :

$s + un mot + $reffew avec mention de la colonne obligatoire :

$regexRenvoiSansEtymon2 := `/$s(_$mot)?_?$vol[,_]+$pagecol/`

$s + un mot appartenant à une liste définie + $reffew :

$regexRenvoiSansEtymon3 := `/$s_$motconnu[,_]?$reffew/`

Si l'on balise les références bibliographiques avant les renvois, on y ajoutera l'interdiction de prendre en compte les expressions $s + référence bibliographique + $reffew. Cela signifie, en pratique, qu'il ne faudra surtout pas effacer ni rendre invisibles les références bibliographiques, mais les garder bien visibles pour cette étape.

supra/infra + page

L'utilisation du marqueur "supra/infra" empêche l'activation du marqueur "hier/ici".

Entre le marqueur ($supra) et la page ($page) peuvent se trouver une virgule et un espace. L'expression régulière est donc `/$supra[,_]*($page)/`.

Cette expression est désignée par la variable $regexRenvoiSansEtymon4.

s./ s.v. + étymon (ou l'inverse : étymon + s./s.v.)

L'étymon peut être précédé du marqueur $s ou du marqueur $sv. Entre le marqueur $s et l'étymon peuvent apparaître un mot tel que *noch*, le marqueur $hier ou encore une référence $reffew. Les combinaisons maximales sont alors les suivantes :

$regexRenvoiAvantEtymon1 := `/$s_($mot_)?($hier[,_]?$reffew)?/` + étymon

$regexRenvoiAvantEtymon2 := `/($sv)/` + étymon

Le marqueur peut parfois suivre l'étymon au lieu de le précéder. Les combinaisons possibles sont les suivantes :

$regexRenvoiAprèsEtymon1 := étymon + `/($s_($mot_)?$reffew/`

$regexRenvoiAprèsEtymon2 := étymon + `/($s_($mot_)?$hier[,_]?$reffew/`

hier/ici + étymon (ou l'inverse : étymon + hier/ici)

$regexRenvoiAvantEtymon3 := `/($s_?)?$hier[,_]?$reffew?($sv)?/`

supra/infra + étymon

$regexRenvoiAvantEtymon4 := `/($supra)[,_]*$page?/`

Extension de l'étymon

Dans les combinaisons activant la mention de l'étymon-cible, des éléments divers peuvent suivre l'étymon dont il est question ci-dessus et préciser la cible. Il peut s'agir d'un marquage alphanumérique (« II 1 » dans « ici 1, 341b, BESTIA II 1 », FEW 25, 961a) ou d'un numéro de note, exprimé par "n" ou "note" suivi d'un numéro (« S. hier 5, 465 LŪRA n 1. », FEW 16, 494a[11]).

Le marquage et le numéro de note sont des informations qui précisent la cible (endroit où l'on renvoie le lecteur). Ils peuvent tous deux être exprimés au moyen d'expressions régulières. Le marquage est représenté par l'expression régulière suivante :

```
/((I|II|III|IV|V|VI|VII|VIII|IX|X)_?
([1-9]_?[0-9]?)?_?[a-z]?_? [α-ω]?_?([a-z]')?_?([α-ω]')?)|
(([1-9]_?[0-9]?)_?[a-z]?_?[α-ω]?_?([a-z]')?_?([α-ω]')?)|
([a-z]_?[α-ω]?_?([a-z]')?_?([α-ω]')?)/
```

Quant au numéro de note, il s'exprime de la façon suivante :

```
/[(n\.|note)_?[1-9][0-9]*/
```

La reconnaissance et le balisage du marquage après un étymon de renvoi sont importants car ils permettent d'éviter que ce marquage soit considéré comme interne à l'article (cf. *tag-numbering*).

5.4.6.6 Renvois multiples

Indicateurs

Les renvois multiples font un usage important de l'implicite : les marqueurs ne sont pas toujours répétés. Comment les reconnaître ? Le premier renvoi de la chaîne aura normalement été reconnu grâce aux règles édictées ci-dessus. Lorsqu'un renvoi est reconnu, il s'agit donc de regarder à sa droite si d'autres constituants peuvent former un nouveau renvoi.

Puisqu'il y a implicite, les combinaisons pour ces renvois supplémentaires sont différentes des précédentes. Il peut s'agir

- d'un étymon seul ;
- d'un étymon + $reffew ;
- d'un étymon + $hier + $reffew ;
- de $reffew ;
- de $hier + $reffew.

Mis à part l'étymon seul, ces combinaisons peuvent s'exprimer en une seule expression régulière `/($hier?) $reffew/`, qui apparaît soit seule, soit après un étymon. Cette expression est appelée $regexRenvoiMultiple.

[11] Cet article est un article de renvoi. Nous n'avons pas trouvé d'exemple de numéro de note dans un renvoi qui apparaîtrait hors articles de renvoi, mais nous prenons malgré tout ce cas en compte, car il nous semble que le renvoi à une note pourrait se justifier aussi en dehors des articles de renvoi.

Incohérences

Un problème sérieux pour le balisage se pose en cas de renvoi multiple lorsque les constituants de deux renvois distincts sont imbriqués :

> « S. noch JOCUS, GRÎS, hier 4, 44a ; 16, 81. » (FEW 20, 12b)

Même manuellement, ce cas est impossible à baliser correctement selon notre modèle. Les constituants devront être balisés séparément, ce qui donnera quatre éléments de renvoi. Dans le traitement informatique, les étymons ne seront pas reconnus comme renvois, puisque le premier ne répond à aucune de nos règles. Les références FEW seront, elles, reconnues et balisées, ce qui est suffisant pour permettre un lien hypertextuel. Les étymons étant de toute façon déjà balisés en tant qu'étymons, l'information n'est pas totalement perdue pour un utilisateur qui voudrait savoir si l'étymon-vedette d'un article fait l'objet de renvois ailleurs dans le FEW : il lui suffira de faire une recherche sur cet étymon dans le FEW.

5.4.6.7 Ordre de détection des renvois

Les expressions régulières ayant été définies, la phase suivante consiste à décider de l'ordre dans lequel elles doivent être détectées. La reconnaissance des renvois avec étymons doit précéder celle des renvois sans étymon, puisque les premiers sont susceptibles de contenir des composants présents dans les seconds. Le paragraphe sera donc traité en deux phases de balisage successives. La détection des renvois multiples doit, quant à elle, se faire chaque fois que l'algorithme trouve un renvoi simple. Pour que la détection des renvois multiples fonctionne, le texte fewien doit absolument être examiné de gauche à droite, depuis le début du paragraphe : cette règle permet, à chaque détection d'un renvoi, d'être certain que les renvois (sans étymon ou avec étymon, selon la phase concernée) présents à sa gauche ont déjà été reconnus.

La détection des renvois s'effectue dans des chaînes virtuelles qui rendent transparentes les balises <b>, <i>, <sc>, <e> et qui rendent visibles les balises <etymon> et </renvoi>. Ces propriétés permettent d'assurer que chaque chaîne virtuelle commence après le renvoi ou l'étymon précédemment traité dans la phase de balisage.

5.4.6.8 Résumé

Dépendances

Tag-renvoi dépend de *tag-etymon* uniquement.

Il utilise également la liste de mots-clés *few-renvoi-base*.

Description

Tag-renvoi identifie et balise les expressions de renvoi qui se trouvent dans un article du FEW.

L'algorithme se compose de trois phases. Il traite d'abord l'entrée, afin de déterminer s'il s'agit d'un article de renvoi (auquel cas les deux phases suivantes ne sont pas

appliquées). La deuxième phase consiste à reconnaître les expressions de renvoi contenant un élément <etymon>, la troisième celles n'activant pas l'élément <etymon>. La deuxième et la troisième phases traitent l'article paragraphe par paragraphe, en incluant les paragraphes de note.

Le traitement de l'entrée consiste à relever entre <entry> et </entry> tous les étymons (préalablement balisés par *tag-etymon*) et à regarder si le premier étymon est un mot-clé de *few-renvoi-base*. Le cas échéant, quelques vérifications sont effectuées quant au nombre de paragraphes et à la présence d'étymons. Si les vérifications sont positives, l'étymon de renvoi est balisé (en incluant toute l'expression de renvoi, depuis le « S./s. » initial jusqu'à la fin de l'entrée). L'article reçoit un attribut *type* avec "renvoi" pour valeur et les deux phases suivantes ne sont pas appliquées.

La reconnaissance des renvois contenant une mention d'étymon s'effectue paragraphe par paragraphe. Les étymons sont détectés grâce à leur balisage et traités dans l'ordre de leur apparition dans le paragraphe. Le contexte gauche et droit de chaque étymon est d'abord défini. Plusieurs expressions régulières (regexRenvoiAvantEtymon1, 2, 3, 4) sont ensuite recherchées dans le contexte gauche, après création d'une chaîne virtuelle rendant transparentes plusieurs balises. Si l'une de ces expressions régulières est détectée, l'algorithme conclut à la présence d'un renvoi, qu'il balise en incluant le marquage alphanumérique ou le numéro de note qui se trouve éventuellement dans le contexte droit de l'étymon. Le marquage qui suit l'étymon est en outre balisé <rpref>. Si aucune des expressions régulières n'a été détectée dans le contexte gauche, le contexte droit de l'étymon est examiné. Plusieurs expressions régulières (regexRenvoiApresEtymon 1, 2) y sont recherchées après création d'une chaîne virtuelle rendant invisibles certaines balises. Si l'une de ces expressions est détectée, l'algorithme conclut à la présence d'un renvoi, qu'il balise depuis l'étymon jusqu'à la fin du texte correspondant à l'expression régulière.

Lorqu'un renvoi a été balisé, l'algorithme vérifie s'il n'est pas suivi d'autres mentions qui formeraient avec lui un renvoi multiple. Ces mentions de renvoi sont détectées au moyen d'une expression régulière (regexRenvoiMultiple, → 5.4.6.6) et balisées individuellement.

La troisième phase de l'algorithme (reconnaissance des renvois ne mentionnant pas d'étymon) s'effectue de la même façon que la deuxième, paragraphe par paragraphe. Pour chaque paragraphe, une chaîne virtuelle est créée dans laquelle les renvois sont visibles, de façon à ne pas traiter les renvois déjà balisés lors de la deuxième phase. Trois expressions régulières (RenvoiSansEtymon) sont recherchées successivement dans le texte. Chaque fois qu'une de ces trois expressions est détectée, la section de texte concernée est balisée et retirée de la section de texte à examiner. Le traitement du paragraphe s'arrête lorsque plus aucune des trois expressions n'est détectée.

5.4.7 Balisage des signatures : tag-signature

Mot d'emprunt. Lt. <sc>genitivus</sc> coexiste avec la forme primitive <sc>genetivus</sc> dp. l'époque classique. Les sens 1 et 2 appartiennent au lt. — Zumthor.	Mot d'emprunt. Lt. <sc>genitivus</sc> coexiste avec la forme primitive <sc>genetivus</sc> dp. l'époque classique. Les sens 1 et 2 appartiennent au lt. — <signature author="Zumthor" lang="french">Zumthor</signature>.

(FEW 4, 102b, GENITIVUS)

5.4.7.1 Objectifs de l'algorithme

L'algorithme *tag-signature* a pour objectif de détecter et de baliser dans le FEW les signatures d'articles ou de parties d'articles (paragraphes ou notes dont le responsable diffère du rédacteur principal de l'article, → 3.7.3).

Le balisage de la signature d'article présente une grande importance pour d'autres algorithmes. L'identification du rédacteur de l'article permet notamment de connaître la langue dans laquelle est rédigé le commentaire final, ce qui constitue un indice pour distinguer ce dernier du champ de la documentation (cf. *split-doc-com*).

Ces raisons justifient que *tag-signature* rétablisse les signatures d'articles là où elles sont implicites et qu'il indique ainsi la métalangue (français ou allemand) utilisée par le rédacteur. La balise <signature> comporte à cet effet obligatoirement deux attributs, *author* et *lang*. Le premier a pour valeurs possibles les noms des rédacteurs du FEW. L'attribut *lang* a pour valeurs possibles, soit "french", soit "german". La signature Zumthor, par exemple, sera balisée comme suit :

<signature author="Zumthor" lang="french"> Zumthor </signature>.

En cas de signature implicite, une balise vide sera insérée à la fin de l'article, de la façon suivante (pour un article rédigé par von Wartburg) :

<signature author="Wartburg" lang="german"/>

5.4.7.2 Détection des signatures explicites

Indicateurs textuels

> ITe. Les divers rédacteurs des volumes 1 à 25 du FEW sont connus : ils sont répertoriés dans les préfaces des différents volumes.

Il est donc envisageable d'en constituer la liste, qui aura la caractéristique d'être finie[12], afin qu'elle serve de critère de détection, à condition que cette liste soit exhaus-

[12] La liste est finie si l'on considère les 25 volumes imprimés du FLa liste est finie si l'on considère les 25 volumes imprimés du FEW ; si l'on prend en compte la refonte en cours et la possibilité d'un FEW évolutif, il est évident que cette liste devient ouverte.

tive et corresponde exactement à ce qu'on trouve dans le FEW.

> Lorsqu'ils ont la fonction de signature, les noms des rédacteurs se déclinent sous des formes variées. Outre le cas majoritaire, le nom de famille complet en minuscules romaines commençant par une majuscule ("Wartburg"), on trouve des abréviations ("Wbg"). Pour les femmes, mais aussi en cas d'ambiguïté ("Colón" pour Germàn Colón – ainsi FEW 14, 641a, VŬLCĀNUS – et "AColón" (à vér.) pour Andres Colón), le nom est précédé de l'initiale du prénom en majuscule, parfois avec une petite espace intermédiaire ("M Hoffert").

Cette liste ne peut se résumer aux noms de famille des rédacteurs, mais doit comprendre toutes les formes sous lesquelles le FEW les mentionne en tant que signataires. Pour reconnaître le rédacteur (= "lemme") qui se cache sous chaque forme, nous utilisons un lemmatiseur très simple, puisque chaque forme est univoque et correspond à un seul rédacteur possible. La liste des formes et des lemmes correspondants est dénommée *few-signature-base*. Elle comporte 84 formes, correspondant à 51 rédacteurs.

> Les noms des rédacteurs ne sont pas toujours des signatures. Il arrive par exemple que le nom d'un rédacteur du FEW soit identique à une référence bibliographique (cf. *tag-biblio*). Le nom d'un rédacteur du FEW peut aussi apparaître, sans qu'il s'agisse d'une référence bibliographique, pour signaler une divergence d'opinion face à la proposition étymologique donnée par le rédacteur de l'article ; Büchi 1996, 161 parle dans ce cas de polyphonie.

Les indicateurs textuels ne sont donc pas spécifiques : ils conduisent à des faux négatifs.

Indicateurs positionnels

> IPo. La signature d'articles entiers constitue la dernière partie du champ alloué au commentaire ; elle est délimitée à droite, soit par le champ alloué aux notes de fin d'article, soit par l'interligne qui sépare l'article en question de l'article suivant.

Le balisage de la signature ne nécessite pas d'avoir préalablement balisé le champ du commentaire. La reconnaissance préalable du champ des notes est suffisant.

> La signature est délimitée à droite par le point final du commentaire qui, en l'absence de notes, sera en même temps le point final de l'article.
> À sa gauche, la signature est précédée, sauf erreur, d'un tiret (semi-)cadratin et d'une espace (facultative).

On peut donc détecter une signature au moyen des signes de ponctuation qui l'entourent.

Ces règles contextuelles simples deviennent plus complexes en cas de signature multiple (articles signés par plusieurs rédacteurs). Les noms des signataires se suivent

alors, séparés par un point-virgule, une virgule ou un point (avec espace éventuelle). Le début et la fin de la séquence respectent toutefois les règles énoncées pour les signatures uniques.

On peut détecter une signature multiple en opérant de droite à gauche à partir de la fin du paragraphe et en vérifiant la ponctuation présente entre chaque item.

Les indicateurs positionnels seuls ne sont pas spécifiques : une séquence de caractères située entre un tiret (semi-)cadratin et le point final de l'article n'est pas toujours une signature.

> Dans cette position et dans le même contexte peuvent apparaître d'autres informations : une référence bibiographique (par ex. "ML 3744" s.v. gĕrmen, -ine) ou un commentaire se résumant à une phrase nominale à un élément (par ex. "Lehnwort" s.v. Ischuria) ou encore un renvoi à un autre article (par ex. "S. noch *TRAGŬLARE" s.v. STRAGŬLUM).

Le balisage préalable des références bibliographiques, des commentaires et des renvois serait nécessaire pour lever ces ambiguïtés.

Les indicateurs positionnels seuls sont, en revanche, fiables si l'on considère la grammaire du FEW. Toutefois, des erreurs ne sont pas à exclure :

> Le tiret (semi-)cadratin situé à gauche de la (séquence de) signature peut être absent (erreur du FEW) ou transcrit comme un trait d'union (erreur lors de la saisie).

En cas d'absence ou de mauvaise transcription de ce tiret (semi-)cadratin, la signature ne serait pas reconnue.

Combinaison des indicateurs

La combinaison des indicateurs textuels et des indicateurs positionnels permet d'énoncer une règle de reconnaissance : un morceau de texte est une signature s'il correspond à une forme de la liste des rédacteurs et si son contexte vérifie les critères positionnels énoncés ci-dessus. La combinaison de ces critères permet de les rendre plus souples et de ne pas devoir vérifier, en particulier, la présence d'un tiret (semi-)cadratin à gauche du nom du rédacteur : seul le contexte droit reste primordial pour reconnaître une signature.

L'algorithme consiste donc à détecter les mots-clés de *few-signature-base* et à vérifier qu'ils sont suivis d'un point et d'une fin de paragraphe. Le contexte gauche est alors pris en compte, uniquement pour la détection des signatures multiples : l'algorithme regarde à gauche du mot-clé s'il est précédé d'un ou de plusieurs mot(s)-clé(s) respectant également les critères positionnels requis. Le cas échéant, ces mots-clés voisins du mot-clé considéré sont également conservés comme étant des signatures.

5.4.7.3 Insertion des signatures implicites

> La majorité des articles ne contiennent pas de signature en structure de surface. En structure profonde, on peut y distinguer deux groupes, selon que la signature y est réellement absente ou implicite. Les articles de renvois constituent le premier groupe : ils ne sont logiquement jamais signés.

On pourrait considérer que les articles de renvoi, qui traduisent une décision non seulement lexicographique, mais aussi lexicologique, sont signés par le responsable du volume. Néanmoins, pour des raisons pratiques, nous excluons du traitement les articles de renvoi, qui ne recevront donc aucune signature.

> Dans le second groupe, où les signatures sont présentes en structure profonde, la règle décisionnelle est la suivante : un article non signé de la partie étymologisée du FEW a implicitement pour rédacteur W. von Wartburg, excepté s'il se trouve dans le volume 20, 33-52 (étymons slaves), où le rédacteur par défaut est Otto Jänicke, dans le volume 24, 97-384, où les rédacteurs par défaut sont von Wartburg et Jänicke.

Pour connaître le signataire par défaut, il suffit donc de connaître le volume où l'on se trouve et, s'il s'agit du volume 20 ou 24, de connaître la ou les page(s) sur laquelle ou lesquelles s'étend l'article.

> Les matériaux d'origine inconnue ou incertaine font également partie du second groupe. Ils ne sont généralement pas signés explicitement (Büchi 1996, 161). L'histoire de leur publication est assez sinueuse, de sorte qu'il est difficile de déterminer un signataire par défaut pour cette partie du FEW. Voici la liste des responsables successifs de ces volumes et fascicules :
> volume 21 : M. Hoffert, J. Hubschmid, H. Lüdtke et M. Müller ;
> volume 22/1, 1-192 et volume 22/2, 1-96 et 193-217 : M. Hoffert ;
> volume 22/1, 193-288 : M. Hoffert et J.-P. Chambon ;
> volume 22/1, 289-316 : J.-P. Chauveau et F. Lagueunière ;
> Volume 22/2, 97-192 : F. Lagueunière, É. Büchi et A. Thibault ;
> volume 22/2, 193-322 : J.-P. Chauveau, F. Lagueunière et A. Thibault ;
> volume 23 : W. v. Wartburg.

Les articles d'origine inconnue sont exclus du traitement de la même manière que les articles de renvois.

L'algorithme de reconnaissance des signatures implicites est le suivant :

```
IF (vol_id = 20) AND (pg_id = [33-52]) →
    attr_author := "Jänicke"
    ; attr_lang := "german"
[] ((vol_id = 20) AND (pg_id != [33-52])) OR
    (vol_id != 20) →
    attr_author := "Wartburg"
    ; attr_lang := "german"
```

FI

5.4.7.4 Attribution de la métalangue

Dès lors que nous avons établi une liste de mots-clés contenant les noms de tous les rédacteurs, il est facile d'y intégrer la correspondance nom du rédacteur – métalangue utilisée par le rédacteur. La liste *few-signature-base* contient trois colonnes, indiquant (1) la forme, (2) le lemme (nom complet du rédacteur) et (3) la métalangue (*french* ou *german*) correspondant au lemme.

L'algorithme pour l'attribution de la métalangue consiste, dès qu'une signature est détectée, à chercher dans la troisième colonne la métalangue correspondant au mot-clé. Cette métalangue est conservée comme valeur d'un attribut *lang* qui, lors du balisage, est ajouté à la balise <signature>.

5.4.7.5 Problèmes particuliers

Un certain nombre de problèmes particuliers doivent être intégrés à l'algorithme.

Variantes graphiques

La méthode choisie – reconnaissance par mots-clés – impose une stricte correspondance entre les mots-clés de la liste et les formes effectivement présentes dans le FEW. La question se pose donc de savoir si toutes les possibilités de variantes peuvent être ajoutées à *few-signature-base*. Trois problèmes sont à envisager.

> Tout d'abord, la petite espace, souvent présente dans le FEW entre des items étroitement liés (« M Hoffert ») est susceptible de disparition (« MHoffert ») ou de transcription comme une espace normale lors de la saisie (« M Hoffert »).

L'intégration de ces variantes dans la liste s'avère facilement réalisable.

> Un deuxième problème concerne le codage des caractères spéciaux dans le nom du rédacteur Lubomir Smiřický.

L'obligation d'un même codage en UTF-8 dans les listes et dans le FEW, ainsi que la vérification des caractères licites lors de la validation d'un article, constituent une première réponse à ce problème (→ 4.3.3 ; 5.2.2.3). Une réponse complémentaire consiste à intégrer à la liste les variantes "Smiricky", "Smirický" et "Smiřicky", afin que ce mot-clé soit reconnu en cas de mauvaise saisie ou d'incohérence du FEW.

> Un dernier problème, plus complexe, concerne la présence de traits d'union de fin de ligne, qui mènent à des coupures du type « Cham-bon ».

Il se trouve toutefois que le problème général des coupures de mots est résolu par l'algorithme de prétraitement *merge-split-words*. Grâce à cet algorithme, les variantes

comportant des traits d'union internes ne doivent pas être ajoutées à la liste de mots-clés.

Interférences

> Des appels de notes apparaissent parfois après la signature, soit directement (ex. FEW 24, 495b ; 525b), soit après le point final du commentaire (ex. FEW 24, 636a).

Afin de faire face à ce type d'inférence, la recherche des mots-clés et la vérification du contexte se font dans une chaîne virtuelle où les appels de note, préalablement balisés, ont été rendus invisibles.

Inconsistances du FEW

> À chaque rédacteur correspond en principe une et une seule métalangue, excepté dans deux cas. D'une part, Hubschmid, quoique germanophone, a rédigé les articles du volume 25 en français, après en avoir rédigé dans les volumes précédents en allemand. Smiřický a quant à lui rédigé ses articles tantôt en français, tantôt en allemand. La préface du fascicule 142 signale que "les articles de ce fascicule rédigés en allemand ont pour auteur soit Walther von Wartburg [...], soit M. Smiřický. Celui-ci a cependant rédigé quelques articles [...] en français" (Gossen 1981 *in* FEW, fascicule 142, Avis au lecteur ; voir aussi Büchi 1996, 138).

La liste *few-signature-smiricky-base* contient les 15 articles que Smiřický a écrits en français, désignés par l'étymon-vedette. Si l'étymon-vedette de l'article traité appartient à cette liste, l'algorithme désignera le français comme métalangue de l'article.

L'algorithme général doit donc être modifié pour intégrer ces deux exceptions :

```
VAR
    attr_lang, attr_author : string ;
    ved : string ;
    vol_id : integer

|[
vol_id := "extraire de <art> la valeur de l'attribut volume"
; IF (keyword = "Hubschmid") AND (vol_id = "25") →
    attr_lang := "french"
[] keyword = "Smiricky" →
    ved := "extraire l'étymon-vedette de l'article"
    ; attr_lang := "extraire de few-smiricky-base la langue
    correspondant à la valeur de la variable ved"
[] ((keyword = "Hubschmid") AND (vol_id != "25"))
    OR
    ((keyword != "Hubschmid") AND (keyword != "Smiricky")) →
    attr_author = "auteur correspondant à keyword dans
```

```
                   few-signature-base"
    ; attr_lang = "langue correspondant à keyword dans
                   few-signature-base"
FI
]|
```

Erreurs du FEW

> Outre les cas particuliers abordés ci-dessus, il faut envisager la possibilité de véritables erreurs, comme l'absence d'une signature dans des articles non rédigés par von Wartburg dans la majeure partie du FEW (ou par Jänicke, dans la section slave).

Il est impossible de repérer ce type d'erreur : les articles en question seront attribués mécaniquement à von Wartburg ou à Jänicke, respectivement. Ce faisant, la rétroconversion reste fidèle à l'attribution qui serait réalisée par un lecteur des volumes papier. Cependant, nous pouvons signaler une absence de signature lorsqu'elle a lieu dans les parties du FEW où tous les articles sont censés être signés.

> Dans les volumes 24 (à partir du fascicule 142) et 25, l'absence d'une signature est d'office une erreur. C'est le cas de l'article ANIMANS (FEW 24, 594a), qui n'est pas signé[13].

La résolution de telles erreurs est problématique, car elle n'est pas algorithmiquement décidable – ni, d'ailleurs, manuellement, à moins de faire des recherches ciblées en dehors du texte du FEW. L'algorithme doit signaler ces erreurs, afin que l'information puisse être ajoutée manuellement.

5.4.7.6 Résumé

Dépendances

Le balisage de la signature nécessite la reconnaissance préalable des articles de renvoi (qui ne contiennent pas de signature) effectuée par *tag-renvoi*, ainsi que le balisage des étymons-vedette (*tag-etymon*). Enfin, il est nécessaire de rendre invisibles les appels de note, qui devront donc être préalablement reconnus et balisés (cf. appels de note).

Tag-signature utilise également deux listes de mots-clés : *few-signature-base* et *few-signature-smiricky-base*.

Description

L'algorithme *tag-signature* identifie le ou les rédacteur(s) d'un article du FEW, ainsi que la métalangue de l'article. Cet algorithme n'est pas appliqué aux articles de renvoi, ni aux matériaux d'origine inconnue. Dans les articles de la partie étymologique du FEW, il détecte tous les mots-clés de la liste *few-signature-base* présents dans chaque

[13] La paternité de cet article doit probablement être attribuée à J.-P. Chambon, qui a rédigé les articles ANIMA, ANIMARE etc.

mascŭlīnus männlich.
1. Judfr. *mallin* m. „membre viril“ IbnEzra.

FIGURE 5.4 – Guillemets entourant les définitions (FEW 6/1, 424b, MASCŪLĪNUS 1)

paragraphe. L'algorithme vérifie ensuite si les mots-clés détectés remplissent les critères positionnels (contexte), de droite à gauche. La détection des mots-clés et la vérification de leur contexte sont effectuées dans une chaîne virtuelle qui rend invisibles les appels de note (cf. point 4.2). Les mots-clés conservés sont balisés. Les attributs *author* et *lang* sont ajoutés pour chaque mot-clé en considérant les cas particuliers présentés par Hubschmid et Smiřický.

Si aucune signature explicite n'a été trouvée, l'algorithme insère à la fin de l'article une balise vide, contenant les attributs *author* et *lang* qui correspondent au rédacteur implicite de l'article, en fonction du volume et des pages où il se trouve.

5.4.8 Balisage des définitions : tag-def

<p>I. Robertv. <i>djamés</i> „se dit de 2 arbres, 2 noix<lb/> qui ont grandi ensemble de façon à faire bloc“ ;<lb/>	<p>I. Robertv. <i>djamés</i> <def>„se dit de 2 arbres, 2 noix<lb/> qui ont grandi ensemble de façon à faire bloc“</def> ;<lb/>

(FEW 4, 93a, GĔMĬNARE I)

5.4.8.1 Objectifs de l'algorithme

L'algorithme *tag-def* a pour objectif de reconnaître la molécule de l'infrastructure qui concerne les définitions des lexèmes, en incluant les guillemets :

texte<def>"Ici se trouve une définition"</def>texte

Tag-def n'a pas pour objectif de vérifier la bonne imbrication et le bon codage des guillemets, puisque ces vérifications ont été effectuées (et d'éventuelles erreurs corrigées) par un algorithme de prétraitement (cf. *streamline-quotes*).

Le balisage des définitions participe à la reconnaissance des unités minimales de traitement (cf. *tag-unit*).

5.4.8.2 Indicateurs de reconnaissance

> IPo. Les définitions n'apparaissent pas dans le champ de l'entrée, mais partout ailleurs dans l'article : documentation, commentaires et notes.
> ITy. Elles sont entre guillemets. Les guillemets utilisés dans le FEW sont bien distincts selon qu'il sont ouvrants ou fermants. Dans la majorité des articles, le guillemet ouvrant est „ (codage unicode U+201E), tandis que le guillemet fermant est “ (codage unicode U+201C) (cf. figure 5.4).

Les guillemets sont un indicateur suffisant pour reconnaître une définition. Il est donc approprié de parcourir chaque paragraphe du FEW (hors entrée) caractère par caractère pour y repérer les guillemets ouvrants et fermants. L'algorithme de prétraitement *streamline-quotes* ayant été préalablement appliqué, il n'est plus nécessaire de vérifier le codage et l'imbrication des guillemets. Chaque paire de guillemets (ouvrant puis fermant) constituera le début et la fin de la chaîne de caractères à baliser en tant que définition.

5.4.8.3 Incohérences

> ITy. À partir du fascicule 154, c'est-à-dire à partir de la page 689 du volume 25, les guillemets utilisés sont différents : le guillemet " (codage unicode U+201C) devient le guillemet ouvrant, tandis que le guillemet fermant est " (codage unicode U+201D).

Le changement typographique opéré à partir de la page 689 du volume 25, dû au changement d'imprimeur consécutif au transfert du FEW de Bâle à Nancy, implique la nécessité de spécifier à l'algorithme quels guillemets sont attendus en fonction du volume et des pages dans lesquels se trouve l'article.

> Le changement typographique a lieu au milieu de l'article ATRIUM (FEW 25, 687b-691b), qui commence à la fin du fascicule 152 et se termine au début du fascicule 154, dans lequel les guillemets ont été modifiés.

Les guillemets de l'article ATRIUM sont normalisés par l'algorithme de prétraitement *streamline-quotes* : les guillemets apparaissant dans la deuxième partie de l'article (à partir de la page 689) sont corrigés par cet algorithme de façon à respecter la définition des guillemets opérante pour la première partie de l'article (pages 687 et 688).

5.4.8.4 Résumé

Dépendances

L'algorithme *tag-def* dépend uniquement de l'algorithme de prétraitement *streamline-quotes*.

Description

Tag-def identifie et balise, dans chaque paragraphe de l'ensemble de l'article sauf dans l'entrée, les définitions des lexèmes, identifiées comme des chaînes de caractères situées entre une paire de guillemets ouvrant et fermant.

Les guillemets ouvrant et fermant opérant pour l'article sont d'abord définis selon le numéro de volume et éventuellement le numéro de page auquel commence l'article ("défnir guillemets").

Pour chaque paragraphe, une chaîne virtuelle est créée, dans laquelle la majorité des balises sont transparentes (cf. partition). La chaîne ainsi créée est parcourue de

gauche à droite, caractère par caractère. Une machine d'état assortie d'une pile détecte les séquences de paires de guillemets (ouvrant et fermant). Le texte compris entre un guillemet ouvrant et un guillemet fermant est balisé <def>. Des situations anormales (guillemets non conformes ou non attendus) conduisent à des alertes et à un arrêt du traitement pour correction manuelle.

5.4.9 Balisage des sigles bibliographiques : tag-biblio

<i>pronom démonstratif</i> „pronom qui exprime une idée d'indication" (1550, Meigret 67 ; seit Fur 1690) ;	<i>pronom démonstratif</i> „pronom qui exprime une idée d'indication" (1550, <biblio>Meigret 67</biblio> ; seit <biblio>Fur 1690</biblio>) ;

(FEW 9, 444b, PRONOMEN 1)

5.4.9.1 Objectifs de l'algorithme

L'algorithme *tag-biblio* a pour objectif de reconnaître les sigles bibliographiques recensés dans le *Beiheft*, quelle que soit leur position dans la structure de l'article, et de les baliser au moyen de l'élément XML <biblio>.

Le balisage doit en outre intégrer les références qui peuvent suivre certains sigles, par exemple un numéro d'édition, un volume, une page, un numéro de vers etc. Le balisage de ces références est essentiel pour le balisage des datations (cf. *tag-date*), car nous avons choisi de ne pas considérer comme molécule de datation les dates d'édition des ouvrages lexicographiques (→ 3.6.4). *Tag-biblio* a donc pour tâche de reconnaître ces dates d'édition, afin que *tag-date* n'ait plus à les traiter.

Les références bibliographiques sont donc balisées comme suit :

<biblio>Sigle + références</biblio>

5.4.9.2 Critères de détection

Indicateurs positionnels

> IPo/ITy. Lorsqu'elles jouent le rôle de molécule dans une cellule lexicale, les références bibliographiques se trouvent généralement dans les parenthèses de précisions (cf. figure 5.5 ; → 3.6.4).
> Toutefois, les parenthèses sont omises en cas de référence bibliographique unique (cf. figure 5.6).

> Des références bibliographiques peuvent aussi apparaître en dehors d'une cellule lexicale, notamment en fin de commentaire, séparées par un tiret long (cf. figure 5.7).

actic (15. jh., AM 29, 400); fr. *auctif* „qui agit, labo-
rieux, diligent“ Geffr, *actif* (seit Oresme, Gdf),

FIGURE 5.5 – Exemple de références bibliographiques entre parenthèses (FEW 24, 115b, ACTIVUS)

„citer en justice, intenter un procès“, Wallis *achoná*
Gl 1, 114, Hérém. *aksyoná* „taquiner, provoquer“.

FIGURE 5.6 – Exemple de référence bibliographique sans parenthèses (FEW 24, 115a, ACTIO 3)

Indicateurs textuels

> ITe. Les sigles bibliographiques sont répertoriés dans le *Beiheft*.
> Une mise à jour du *Beiheft* a été effectuée en 2010 pour prendre en compte les nombreux sigles nouveaux ajoutés au FEW depuis la parution du supplément au *Beiheft* en 1989. On ne peut être absolument certain de l'exhaustivité de cette liste, mais les oublis sont probablement peu nombreux.

Combinaison des indicateurs

Les indicateurs positionnels évoqués ci-dessus ne sont pas assez spécifiques pour reconnaître de façon fiable une référence bibliographique. Les critères textuels sont dès lors indispensables. La mise à jour du *Beiheft* permet de disposer d'une liste qui peut être utilisée par l'algorithme de détection. Cette liste, comprenant tous les sigles bibliographiques répertoriés dans le *Beiheft* mis à jour, a été établie sous le nom de *few-bibbase*.

Incohérences du FEW

> Un problème critique est l'existence dans le FEW d'incohérences dans les références bibliographiques. Les abréviations bibliographiques contenant deux parties logiques, telles que *BullAin* (pour *Bulletin de la Société des naturalistes et et des archéologues de l'Ain*, Bourg-en-Bresse 1896—), apparaissent tantôt en un seul mot, tantôt avec une (semi-)espace entre les deux parties (*Bull Ain*).

Pour résoudre ce problème d'espaces facultatifs, l'algorithme de détection de mots-clés ajoute à chaque mot-clé une variante « espacée », dans laquelle une espace a été

Es ist sodann wiederholt und in verschiedenen
milieus entlehnt worden, doch ohne dass es sich
richtig verwurzelt hat (II 1). Ganz vereinzelt
auch CONGEMINATIO „verdoppelung“ entlehnt [2])
(2). — Haust Et 70.

FIGURE 5.7 – Exemple de référence bibliographique en fin de commentaire (FEW 4, 93a, GĔMĪNARE)

ajoutée devant chaque majuscule interne.

> Certaines sources se présentent sous des sigles bibliogroahiques divers : par exemple, la référence à l'ouvrage d'A. Darmesteter et D.-S. Blondheim *Les gloses françaises dans les commentaires talmudiques de Raschi* (Paris, 1929) se fait normalement sous le sigle *Rs*, mais elle apparaît aussi (de façon erronée) sous la forme *Raschi* (FEW 6/1, 5a, MACELLUM).
> Par ailleurs, le *Beiheft* ne répertorie pas toutes les références bibliographiques du FEW : les références citées sous la forme du nom complet de l'auteur comme « Diderot », qui ne constituent pas des sigles à proprement parler, n'apparaissent pas dans le *Beiheft*, pour la simple raison qu'il ne s'agit pas de sources du FEW, mais de sources primaires, citées comme telles.

L'utilisation des seuls critères textuels (détection de mots-clés dans une liste de sigles établie à partir du *Beiheft*) n'est donc pas totalement fiable. Les références bibliographiques non présentes dans la liste, soit à cause de la non exhaustivité du *Beiheft*, soit à cause d'incohérences du FEW, ne seront pas repérées. Ce problème est résolu en partie par un algorithme de post-traitement (cf. *show-untagged-text*).

5.4.9.3 Résolution des collisions

La liste *few-bib-base* comporte des cas de collisions, c'est-à-dire des items identiques à d'autres items appartenant à d'autres listes. Les listes concernées sont *few-concept-base* (1 collision), *few-etymon-base* (53 collisions), *few-geoling-base* (17 collisions), *few-numbering-roman-numbers-base* (1 collision), *few-signature-base* (17 collisions) et *few-signature-smiricky-base* (1 collision). Par exemple, Bacon renvoie dans le FEW soit à un étymon, soit à un sigle bibliographique : il appartient donc aux deux listes *few-etymon-base* et *few-bib-base*.

Ces collisions posent problème si on ne les résout pas, car les items en question risquent d'être balisés deux fois différemment (→ 5.2.2.3), comme ceci :

<etymon><biblio>Bacon</biblio></etymon>.

La collision avec few-concept-base est résolue en n'appliquant pas *tag-biblio* dans le champ de l'entrée. Les collisions avec les listes *few-etymon-base*, *few-signature-base* et *few-signature-smiricky-base* sont résolues par le séquençage des algorithmes (*tag-etymon* et *tag-signature* sont appliqués avant *tag-biblio*) et par la création de chaînes virtuelles dans lesquelles les balises <etymon> et <signature> sont rendues invisibles avec leur contenu (→ 5.2.2.3). La collision avec *few-numbering-roman-numbers-base* est résolue par *tag-numbering*, qui supprime les éléments <biblio> apparaissant au milieu d'un marqueur alphanumérique (cf. *tag-numbering*).

Les collisions avec la liste *few-geoling-base* posent problème, car elles ne peuvent être résolues automatiquement par les algorithmes *tag-biblio* et *tag-geoling*. Elles sont résolvables dans certains cas par l'algorithme *tag-unit*. Les autres cas requièrent une inspection et une prise de décision par un expert humain. Les items appartenant à ces deux listes sont donc marqués comme ambigus pour traitement ultérieur, de la façon suivante. Le premier des deux algorithmes, *tag-biblio*, les balise avec un attribut *status*

indiquant la possibilité qu'il s'agisse d'un sigle bibliographique, comme ceci :

<biblio status="geoling ?">Montaigne</biblio>

Les collisions avec la liste *few-geoling-base* sont en outre insensibles à la présence d'un point à la fin des mots-clés, que ces derniers proviennent de *few-bib-base* ou de *few-geoling-base*, pour remédier à deux problèmes différents : tout d'abord, la possibilité qu'un mot-clé (essentiellement bibliographique) soit suivi d'un point de par sa position dans la syntaxe fewienne ; ensuite, l'oubli (provenant du FEW lui-même ou de la saisie) d'un point appartenant à un mot-clé (essentiellement géolinguistique) dans le texte du FEW.

5.4.9.4 Détection des références suivant le sigle bibliographique

> Les références bibliographiques peuvent être accompagnées de références précises indiquant le numéro d'édition, le volume, la page, le vers etc.

Les références attenantes au sigle bibliographique doivent être intégrées dans l'élément XML <biblio>. Elles peuvent être reconnues au moyen des expressions régulières que voici, l'expression \$number remplaçant la séquence ([1-9][0-9]*) qui désigne un nombre :

```
/_?$number(,?_$number(_?[----]_?$number)?)*/
```

Cette expression régulière reconnaît des références telles que "Gdf 10, 704" ou "Gdf 10, 704-705". Nous la nommerons \$pat1.

```
/_?$number?(_?p_?$number)?(,_?$number)*/
```

Cette expression, que nous nommerons \$pat2, reconnaît des références telles que "ALF 865 p 991", "ALEC 987* p 97" ou encore "ALF 865 p 991, 888, 670".

```
/_?$number(_?[----]_?$number)?/
```

Cette expression, que nous nommerons \$pat3, reconnaît des références telles que "ALF 311-313" ou "Trév 1704-1771".

Afin de baliser également des successions de références telles que "RlFn 1, 79 ; 7, 175–7, 212" (FEW 16, 176a, HASE), nous pouvons écrire l'expression régulière suivante, qui reconnaît après le sigle bibliographique une des trois expressions décrites ci-dessus ([(\$pat1)(\$pat2)(\$pat3)]) et y ajoute la possibilité que l'expression reconnue soit suivie, après un point-virgule, de nouvelles expressions parmi les trois décrites :

```
/([($pat1)($pat2)($pat3)](_?;_?[($pat1)($pat2)($pat3)])*)?/
```

5.4.9.5 Cas particulier : les dates d'édition

L'expression régulière \$number, telle qu'elle est définie ci-dessus, peut détecter une datation exprimée en année (« 1869 »), ce qui semble problématique dans la mesure où les dates sont censées être balisées en tant que molécules de datation par l'algorithme *tag-date*, et non en tant que références bibliographiques. Toutefois, dans le FEW, une référence bibliographique n'est jamais (à notre connaissance) suivie directement

(c'est-à-dire sans ponctuation intermédiaire) d'une date qui ne soit pas une date d'édition. La date d'édition qui accompagne les sources lexicographiques telle que « 1771 » dans "Trév 1771" pose un problème de modélisation qui a été discuté et résolu ailleurs (→ 3.6.4) : le modèle que nous avons élaboré ne balise pas ces dates d'édition comme des molécules de datation, mais les rattache à la référence bibliographique. Cette particularité nécessite que *tag-biblio* soit appliqué avant *tag-date*.

> ITe. Une liste des sources lexicographiques, comprenant les dates d'édition, est fournie par le *Beiheft*.

Ces dates d'édition étant connues, il est possible de les intégrer directement dans *few-bib-base*, à la suite du nom du dictionnaire. L'ensemble (source + date(s) d'édition(s)) sera dès lors reconnu et balisé comme une seule référence bibliographique.

Il est toutefois à remarquer que l'expression régulière définie ci-dessus aurait de toute façon repéré les dates d'édition et les aurait incluses dans la référence. L'intérêt principal d'une intégration de ces dates dans la liste *few-bib-base* est de pouvoir par la suite vérifier qu'une date qui suit un sigle bibliographique lui correspond, notamment dans le cas de fourchettes telles que Trév 1704-1771, balisée <biblio>Trév 1704-1771</biblio> dans un premier temps. Si, pour des raisons d'exploitation du FEW, l'on décide de séparer cet élément bibliographique en deux références (<biblio>Trév 1704</biblio> et <biblio>Trév 1771</biblio>), la présence de Trév 1771 dans la liste des sigles confirmera la validité de ce balisage. Au contraire, si la date associée à une référence bibliographique ne correspondait pas à une date d'édition répertoriée dans la liste, la date devrait être vérifiée et, éventuellement, identifiée comme une molécule de datation à part entière et non comme un élément d'une référence bibliographique.

5.4.9.6 Résumé

Dépendances

L'algorithme *tag-biblio* dépend des algorithmes *tag-def*, *tag-etymon*, *tag-signature* et *tag-appelnote*. Les renvois et les langues d'étymon sont également rendus invisibles dans la chaîne virtuelle pour faciliter l'exécution de l'algorithme, mais l'application de *tag-renvoi* et de *tag-lang-etymon* n'étant pas obligatoire pour le bon fonctionnement de *tag-biblio*, ces deux algorithmes ne rentrent donc pas dans la liste des dépendances à proprement parler.

Tag-biblio utilise la liste de mots-clés *few-bib-base* contenant les sigles bibliographiques actuellement répertoriés par le *Beiheft*.

Description

L'algorithme *tag-biblio* identifie et balise les références bibliographiques qui sont contenues dans la liste *few-bib-base*. Il n'opère pas sur le champ de l'entrée, mais sur tout le reste de l'article, paragraphe par paragraphe, y compris sur les paragraphes de notes.

Pour chaque paragraphe, une chaîne virtuelle est créée dans laquelle toutes les balises sont transparentes, excepté <appelnote>, <e>, <def>, <geoling>, <i>, <lang_-etymon>, <renvoi>, <signature>, qui sont invisibles avec leur contenu (cf. partition).

La recherche des mots-clés provenant de la liste *few-bib-base* est effectuée dans cette chaîne virtuelle.

Le balisage inclut les références qui se trouvent éventuellement à la droite du mot-clé. Celles-ci sont détectées au moyen d'une expression régulière ("inclure références attenantes").

Chaque mot-clé balisé reçoit un attribut *status* dont la valeur peut être "ok" ou "geoling ?". Cette dernière valeur est attribuée si le mot-clé de *few-bib-base* rentre en collision avec un mot-clé de *few-geoling-base* ou *few-geoling-error-base*. Deux mots-clés sont considérés comme faisant l'objet d'une collision s'ils sont identiques après suppression des points finals éventuels. Dans le cas d'une collision, la valeur "geoling ?" indique l'ambiguïté, et un avertissement est émis.

5.4.10 Balisage des sigles géolinguistiques : tag-geoling

I. 1. Apr. <i>encar</i> „commencer“ (périg. Marche), <i>enquar</i>, <i>encoar</i> (Avignon 1390, Pans)	I. 1. <geoling status="ok">Apr.</geoling> <i>encar</i> „commencer“ (<geoling status="ok">périg.</geoling> <geoling status="ok">Marche</geoling>), <i>enquar</i>, <i>encoar</i> (<geoling status="ok">Avignon</geoling> 1390, Pans)

(FEW 4, 622b, ĬNCHOARE)

5.4.10.1 Objectifs de l'algorithme

L'algorithme *tag-geoling* a pour objectif de reconnaître les sigles géolinguistiques, c'est-à-dire les termes techniques du lexique fewien qui apparaissent dans les molécules de l'étiquette géolinguistique et de la localisation. Ces deux molécules sont regroupées en un seul élément XML, car, malgré leur rôle différent, elles peuvent se réaliser au moyen des mêmes termes techniques du lexique fewien (→ 3.6.4). Leur nature est semblable, même si leur fonction est différente. C'est la position du terme dans la syntaxe fewienne qui détermine s'il s'agit d'une étiquette géolinguistique ou d'une localisation.

Tag-geoling doit en revanche opérer une distinction entre les étiquettes géolinguistiques galloromanes (apparaissant essentiellement dans la documentation, par exemple *afr.* ou *dauph.*) et les étiquettes géolinguistiques non galloromanes (apparaissant essentiellement dans le commentaire et les notes des articles, par exemple *it.* ou *roum.*). Les premières sont balisées <geoling>, tandis que les secondes sont balisées <lang> (pour *langue*, conformément aux standards de balisage XML-TEI).

Outre les raisons dues à la modélisation du FEW (→ 3.6.3, le balisage des sigles géolinguistiques est essentiel à la reconnaissance des cellules lexicales (cf. *tag-unit* et *tag-precisions*). La distinction entre les étiquettes galloromanes et les étiquettes non galloromanes se justifie également pour des raisons algorithmiques, notamment pour la bonne application de l'algorithme *split-doc-com* qui sépare les parties documentation et commentaire d'un article.

Tag-geoling reconnaît uniquement les termes qui font partie de la liste du *Beiheft* (dans sa version mise à jour par Jean-Paul Chauveau et Yan Greub en 2010).

L'algorithme doit en outre corriger les erreurs éventuelles du FEW (sigles erronés, connus depuis la constitution de l'index du FEW, → 1.2). Il doit également signaler, pour traitement ultérieur et vérification manuelle, les cas ambigus où un sigle géolinguistique pourrait aussi être un sigle bibliographique.

Les sigles géolinguistiques galloromans sont balisés comme suit :

<geoling status="ok">Sigle</geoling>

tandis que les étiquettes non galloromanes sont balisées ainsi :

<lang status="ok">Sigle</lang>.

5.4.10.2 Indicateurs de reconnaissance

> IPo. Lorsqu'ils ont la fonction d'étiquette géolinguistique, les sigles géolinguistiques se trouvent au tout début de l'unité minimale de traitement. Ils peuvent cependant être ellipsés selon la règle générale (et, donc, s'avérer absents en structure de surface). Lorsqu'ils jouent le rôle de localisation, ils apparaissent dans les parenthèses de précisions.
> ITe. Les étiquettes géolinguistiques sont en outre répertoriées dans le *Beiheft*.

Les indicateurs positionnels évoqués ci-dessus ne sont pas assez spécifiques pour reconnaître de façon fiable un sigle géolinguistique. Les critères textuels sont dès lors indispensables. L'existence d'un répertoire des étiquettes géolinguistiques dans le *Beiheft* permet de disposer d'une liste qui pourra être utilisée par l'algorithme de détection. Encore faut-il que cette liste soit exhaustive.

> Une mise à jour du *Beiheft* a été effectuée en 2010 pour prendre en compte les nombreux sigles nouveaux ajoutés au FEW depuis le supplément du *Beiheft* de 1989. On ne peut être absolument certain de l'exhaustivité de cette nouvelle liste, mais les oublis sont probablement très rares.

Une liste *few-geoling-base* a donc pu être établie, comprenant toutes les étiquettes géolinguistiques répertoriées dans le *Beiheft* mis à jour.

> Toutes les étiquettes géolinguistiques et localisations ne sont pas répertoriées dans le *Beiheft*. Par exemple, *Belgique* (FEW 10, 149b, RECIPROCUS 1 a), considéré sans doute comme d'interprétation évidente, ne s'y trouve pas. Il ne s'agit pas là d'un oubli : tout nom de lieu peut être donné comme sigle géolinguistique, même s'il n'appartient pas à la liste des sigles, à condition qu'une référence bibliographique l'accompagne.

L'utilisation des critères textuels (détection de mots-clés dans une liste de sigles établie à partir du *Beiheft*) permet de reconnaître uniquement les termes techniques répertoriés. Les étiquettes et localisations qui ne font pas partie de la liste *few-geoling-base* ne seront pas détectées par l'algorithme. Ce problème est résolu en partie par un

algorithme de post-traitement (cf. *show-untagged-text*).

5.4.10.3 Distinction des étiquettes galloromanes et non galloromanes

Les étiquettes galloromanes sont balisées <geoling>, tandis que les autres sont balisées <lang>. La liste *few-geoling-base* indique pour chaque étiquette, dans la deuxième colonne, à quel groupe elle appartient. Chaque mot-clé de la liste détecté par l'algorithme est balisé <geoling> ou <lang> en fonction de cette indication.

5.4.10.4 Erreurs du FEW

> Un problème critique est l'existence dans le FEW d'erreurs et d'incohérences provenant de l'édition ou de la rédaction. Il arrive en effet qu'une même entité géohistorique soit citée sous la forme de plusieurs sigles géolinguistiques. Par exemple, *Calv.* (abréviation de *Calvados*) se présente aussi sous la variante *Calvad.* (FEW 3, 159a, DRILLEN). Or, la seule forme répertoriée dans le *Beiheft* est *Calv.*

L'utilisation des critères textuels (détection de mots-clés dans une liste de sigles établie à partir du *Beiheft*) permet de reconnaître uniquement les termes techniques connus. Les étiquettes et localisations qui ne sont pas répertoriées dans la liste *few-geoling-base* ne seront pas repérées. Ce problème est résolu en partie par un algorithme de post-traitement (cf. *show-untagged-text*).

> La constitution de l'index partiel des formes du FEW (ATILF 2003) a révélé la présence de 166 étiquettes géolinguistiques erronées. Étant donné que l'Index ne recense pas toutes les formes du FEW, mais seulement une sélection d'entre elles, cette liste de 166 étiquettes constitue un sous-ensemble seulement des étiquettes erronées apparaissant dans le FEW.

Ces sigles ont été rassemblés en une liste *few-geoling-error-base*, qui contient pour chaque sigle erroné sa forme correcte. Lorsqu'un sigle appartenant à cette liste est détecté, il est directement corrigé dans l'article du FEW, puis balisé. Un attribut *status* indique alors, par la valeur "fixed", que le sigle n'est pas le sigle original de la version papier.

5.4.10.5 Résolution des collisions

La liste *few-geoling-base* comporte des cas de collisions, c'est-à-dire des items identiques à d'autres items appartenant à d'autres listes. Les listes concernées sont *few-bib-base* (20 collisions), *few-etymon-base* (159 collisions) et *few-signature-base* (1 collision). Par exemple, *Cassis* renvoie dans le FEW soit à un étymon, soit à un sigle géolinguistique : il appartient donc aux deux listes *few-etymon-base* et *few-geoling-base*.

Ces collisions posent problème si on ne les résout pas, car les items en question

risquent d'être balisés deux fois différemment (→ 5.2.2.3), comme ceci :

<etymon><geoling>Cassis</geoling></etymon>.

Les collisions avec les listes *few-etymon-base* et *few-signature-base* sont résolues par le séquençage des algorithmes (*tag-etymon* et *tag-signature* sont appliqués avant *tag-geoling*) et la création de chaînes virtuelles dans lesquelles les balises <etymon> et <signature> sont rendues invisibles avec leur contenu (→ 5.2.2.3).

Les collisions avec la liste *few-bib-base* posent problème, car elles ne peuvent être résolues automatiquement par les algorithmes *tag-biblio* et *tag-geoling*. Elles sont résolvables dans certains cas par l'algorithme *tag-unit*. Les autres cas requièrent une inspection et une prise de décision par un expert humain. Les items appartenant à ces deux listes sont donc marqués comme ambigus pour traitement ultérieur. Le premier des deux algorithmes appliqués dans la séquence de rétroconversion, *tag-biblio*, les balise avec un attribut *status* indiquant la possibilité qu'il s'agisse d'un sigle géolinguistique, comme ceci :

<biblio status="geoling ?">Montaigne</biblio>

5.4.10.6 Résolution des enchâssements

Problématique

Outre les collisions exactes, l'utilisation de listes de mots-clés peut conduire à des enchâssements, c'est-à-dire des inclusions de mots-clés d'une liste dans les mots-clés d'une autre liste : par exemple, l'étiquette *AT* (« Autriche ») est enchâssée dans le sigle bibliographique *ATrP* ; le sigle bibliographique *Barba* est enchâssé dans le concept *orgue de Barbarie* et dans les étymons *Barbacola*, *Barbara* et *Barbaria*. En réalité, des enchâssements comme ceux-ci ne posent aucun problème, car les algorithmes, lors de la détection de mots-clés, vérifient la présence de délimiteurs (signes de ponctuation ou signes d'espacement, → 4.2.2) autour des mots-clés : *AT* ne sera pas détecté par l'algorithme *tag-geoling* dans *ATrP*, car il n'y est pas suivi d'un délimiteur, mais d'une lettre (*r*). En revanche, l'enchâssement de l'étiquette géolinguistique *Ain* dans le sigle bibliographique *Bull Ain* est problématique, car les caractères d'espacement autour du terme *Ain* incitent l'algorithme *tag-geoling* à le détecter et à le baliser comme une étiquette géolinguistique, alors qu'il s'agit de la partie d'un sigle bibliographique.

Ce type de collision lié à des enchâssements problématiques apparaît fréquemment entre les deux listes *few-bib-base* (utilisée par *tag-biblio*) et *few-geoling-base* (utilisée par *tag-geoling*). La majorité des enchâssements problématiques sont des sigles bibliographiques enchâssant des étiquettes géolinguistiques, mais l'inverse apparaît également. Par ailleurs, certains enchâssements apparaissent entre *few-bib-base* et *few-geoling-error-base*, ce qui pose un problème majeur, puisque les mots-clés de *few-geoling-error-base* sont corrigés (cf. ci-dessus). Par exemple, *Roch* (de *few-geoling-error-base*, corrigé en *Roch.* avec point final) est enchâssé dans *Bull Roch* (de *few-bib-base*) : il ne faudrait pas que *tag-geoling* corrige *Roch* dans ce cas.

Résolution

L'application de *tag-biblio* avant *tag-geoling* dans le séquençage des algorithmes, ainsi que l'invisibilité des balises <biblio> dans *tag-geoling*, résoudrait le problème des cor-

rections de sigles erronés ainsi que la plupart des problèmes d'enchâssements. Toutefois, cette solution ne résout pas le problème des sigles bibliographiques enchâssés dans des étiquettes géolinguistiques : ces dernières, par suite de l'invisibilité des premiers, ne seraient pas reconnues par *tag-geoling*. La seule solution fiable pour résoudre tous les types d'enchâssements consiste à rendre les balises <biblio> transparentes lors de la détection des mots-clés géolinguistiques, puis à examiner (toujours de façon totalement automatique) les inclusions de balises et à supprimer les balises internes. Par exemple, l'enchâssement

<geoling>item1 <biblio>item2</biblio></geoling>

sera résolu en supprimant les balises <biblio>, ce qui donnera <geoling>item1 item2</geoling>.

En cas d'inclusion de balises qui ne représenteraient pas un enchâssement, mais une collision exacte, la règle est différente : ce sont les balises <geoling> (ou <lang>) qui sont automatiquement supprimées. En effet, la balise <biblio> conserve, par son attribut *status*, l'information qu'il s'agit d'une ambiguïté à lever manuellement. Les séquences

<geoling><biblio status = "geoling ?">item</biblio></geoling> et
<biblio status = "geoling ?"><geoling>item</geoling></biblio>

seront donc résolues toutes deux en <biblio status = "geoling ?">item</biblio>.

Remarque. Les mots-clés de *few-geoling-error-base* enchâssés dans une balise <biblio> ne sont ni corrigés ni balisés.

5.4.10.7 Résumé

Dépendances

L'algorithme *tag-geoling* dépend des algorithmes *tag-etymon*, *tag-signature*, *tag-def* et *tag-biblio*.

Par ailleurs, il utilise la liste de mots-clés *few-geoling-base*, qui contient les sigles géolinguistiques répertoriés par le *Beiheft*, ainsi que la liste *few-geoling-error-base*, qui contient les sigles géolinguistiques erronés à corriger.

Description

L'algorithme *tag-geoling* identifie et balise dans un article du FEW les sigles géolinguistiques et les mentions de langue qui sont contenus dans les listes *few-geoling-base* et *few-geoling-error-base*. Il n'opère pas sur le champ de l'entrée, mais bien sur tout le reste de l'article, y compris les paragraphes de note.

Pour chaque paragraphe, deux traitements sont appliqués : d'abord la reconnaissance et le balisage des mots-clés de *few-geoling-base* ; ensuite la reconnaisance, la correction et le balisage des mots-clés de *few-geoling-error-base*. Dans chacun des deux cas, le traitement se fait sur une chaîne virtuelle dans laquelle toutes les balises sont transparentes, excepté <appelnote>, <def>, <geoling>, <i>, <lang_etymon>, <renvoi> et <signature>, qui sont invisibles avec leur contenu (cf. partition). Les balises <biblio> sont transparentes afin de détecter les cas d'enchâssements (cf. plus loin). L'invisibi-

lité de <geoling> et de <lang> est requise pour éviter que les mots-clés balisés lors du premier traitement soient balisés à nouveau lors du second.

Après ces deux traitements, les enchâssements entre les éléments <geoling> (ou <lang>) et <biblio> sont résolus.

Le premier traitement est assez simple. Les mots-clés détectés dans la chaîne virtuelle sont balisés <geoling> ou <lang> (selon les informations données dans *few-geoling-base*) et reçoivent un attribut *status* dont la valeur est "ok".

Le second traitement est plus complexe. Les mots-clés de *few-geoling-error-base* sont détectés dans la chaîne virtuelle. Ceux qui se trouvent dans un élément <biblio> ne sont pas conservés. Les autres sont remplacés dans le texte par le sigle correct, donné dans la seconde colonne de *few-geoling-error-base*. Les sigles ainsi corrigés sont détectés à nouveau pour balisage (cette re-détection est nécessaire, car leur position a pu être légèrement décalée dans la chaîne virtuelle après leur correction et après d'autres corrections effectuées précédemment dans le même paragraphe). Ils reçoivent un attribut *status* dont la valeur est "fixed", indiquant qu'ils sont le résultat d'une correction, et un avertissement est émis.

Les enchâssements entre les éléments <geoling> (ou <lang>) et <biblio> sont résolus en supprimant les balises internes. Les items <geoling> ou <lang> enchâssés dans des <biblio> sont supprimés, de même que les items <biblio> enchâssés dans des <geoling> ou des <lang>. Les collisions, quant à elles, sont résolues par la suppression des balises <geoling> et <lang>, que ces dernières soient internes ou externes.

5.4.11 Séparation de la documentation et du commentaire : split-doc-com

2. Nfr. pronominel adj. „qui est de la nature du pronom, qui appartient au pronom“ (seit 1714, s. Trév 1743), verbe pronominal „réfléchi“ (seit ca. 1720, s. Trév 1743), adjectif pronominal „susceptible de devenir pronom par un léger changement de forme (p. ex. quelque)“ (ca. 1750). - Ablt. Nfr. employer pronominalement „comme verbe pronominal“ (seit Boiste 1829) ; pronominaliser v. a. „donner la forme pronominale“ Lar 1875. 1 entlehnt aus pronomen, 2 aus dem adj. pronominalis „zum fürwort gehörig“.	<doc> 2. Nfr. pronominel adj. „qui est de la nature du pronom, qui appartient au pronom“ (seit 1714, s. Trév 1743), verbe pronominal „réfléchi“ (seit ca. 1720, s. Trév 1743), adjectif pronominal „susceptible de devenir pronom par un léger changement de forme (p. ex. quelque)“ (ca. 1750). - Ablt. Nfr. employer pronominalement „comme verbe pronominal“ (seit Boiste 1829) ; pronominaliser v. a. „donner la forme pronominale“ Lar 1875. </doc> <com> 1 entlehnt aus pronomen, 2 aus dem adj. pronominalis „zum fürwort gehörig“. </com>

(FEW 9, 444b-445a, PRONOMEN)

5.4.11.1 Objectifs de l'algorithme

L'algorithme *split-doc-com* a pour objectif de baliser chacun des paragraphes d'un article (hors articles de renvoi) en indiquant s'il s'agit d'un paragraphe de documentation, de commentaire ou un paragraphe mixte, contenant les deux champs. Pour ce faire, il inclut chaque élément <p> dans un élément XML <doc>, <com> ou <mixt>.

L'algorithme doit également indiquer si l'article respecte le schéma général "champ documentaire suivi d'un commentaire" ou s'il s'écarte de ce schéma. L'article (marqué par <art>...</art>) reçoit un attribut *type* qui peut recevoir plusieurs valeurs : "doc-com" (schéma général respecté), "doc-only" (paragraphe[s] de documentation uniquement), "doc-mixt" (paragraphe[s] de documentation et paragraphe[s] mixte[s]), "mixt-only" (paragraphe[s] mixte[s] uniquement) ou "entry-only" (article ne contenant aucun paragraphe en dehors de l'entrée). Si l'article appartient aux volumes 21 à 23, l'attribut *type* reçoit automatiquement la valeur "doc-concept".

Outre les raisons évoquées pour la modélisation du FEW (→ 3.7), la séparation des deux champs est nécessaire pour le bon fonctionnement d'autres algorithmes de la séquence de rétroconversion tels que *tag-numbering* (balisage des marqueurs alphanumériques), *split-mixt-art* (séparation de deux parties documentation et commentaire au sein d'un même paragraphe) et tag-form (balisage des signifiants des lexèmes).

5.4.11.2 Balisage de chaque paragraphe

Choix du paragraphe comme unité de balisage

> La plupart des articles du FEW respectent le schéma "documentation suivie d'un commentaire". Toutefois, un certain nombre d'articles, essentiellement dans les premiers volumes, mélangent les deux champs (cf. Büchi 1996, 135-136).

L'existence de ces cas atypiques rend caduque un balisage séparé de la documentation et du commentaire en deux champs homogènes. Suivant les décisions prises lors de la construction du modèle (→ 3.7.2), chaque paragraphe de l'article est typé <doc>, <com> ou <mixt> selon son contenu.

5.4.11.3 Indicateurs de reconnaissance

> IPo. Le commentaire constitue généralement un (ensemble de) paragraphe(s) situé(s) à la fin de l'article, avant le champ des notes. Toutefois, tous les articles ne respectent pas ce schéma général : le commentaire peut être totalement absent de l'article ou se trouver inséré au milieu d'un paragraphe de documentation (Büchi 1996, 137).

La position du paragraphe dans l'article ne constitue pas un critère de reconnaissance fiable pour déterminer s'il s'agit d'un paragraphe de documentation ou d'un paragraphe de commentaire. Les critères positionnels s'avérant non fiables, il est nécessaire de trouver des critères textuels (mots-clés ou expressions régulières) ou des

critères structurels (séquences d'éléments déjà balisés) dont la présence à l'intérieur du paragraphe constituerait un indice de sa nature. La question est de déterminer si certains types d'information ou certains éléments textuels apparaissent dans un des deux champs seulement et pourraient dès lors constituer un critère de reconnaissance.

> IS. La totalité des types d'information présents dans la documentation (marquage alphanumérique, molécules de l'unité minimale de traitement etc.) peuvent également apparaître dans le commentaire.

La documentation ne contient aucun élément spécifique sur la présence duquel nous pourrions nous fonder pour distinguer les deux champs.

> ITe. Certains mots reviennent fréquemment dans le commentaire et n'apparaissent jamais (sauf exception tout à fait rare) dans la documentation : il s'agit de termes techniques tels que lehnwort (ou emprunt dans les articles rédigés en français).

Ces termes techniques constituent un critère de détection intéressant, car assez spécifique pour éviter les faux négatifs. Leur présence dans un paragraphe indique avec une haute probabilité l'appartenance de ce paragraphe au commentaire ou du moins l'existence d'une partie de commentaire dans le paragraphe. Deux listes de ces termes ont donc été constituées, l'une en allemand (*few-com-wort-base*) et l'autre en français (few-com-mot-base). Les termes ont été choisis selon une démarche heuristique, en examinant dans la totalité du FEW un maximum de commentaires y compris très courts, de façon à ce que la liste soit valable même pour un commentaire ne comprenant qu'un seul mot comme *Lehnwort*.

Dans le cas d'articles dont la métalangue est le français, un problème se pose : les termes de few-com-mot-base sont susceptibles d'apparaître dans les définitions des lexèmes. Ils constituent dès lors un critère non fiable (menant à de faux positifs). Le balisage préalable des définitions, associé à leur invisibilité lors de la recherche des mots-clés (construction d'une chaîne virtuelle), apporte une solution à ce problème. De même, les termes techniques en gras ou en italiques ne seront pas pris en compte, grâce à l'invisibilité des éléments <b> et <i> dans la chaîne virtuelle.

Les collisions des mots-clés de *few-com-wort-base* et de few-com-mot-base avec les listes *few-bib-base*, *few-geoling-base* et *few-etymon-base* sont résolues également par l'invisibilité des éléments <biblio>, <geoling> et <etymon> préalablement balisés.

Algorithme de balisage

L'algorithme de balisage consiste en définitive à vérifier la présence, au sein du paragraphe, d'un terme technique consigné dans *few-com-wort-base* ou dans *few-com-mot-base*. Le choix de la liste de mots-clés dépend de la métalangue de l'article. Celle-ci est connue grâce à la reconnaissance préalable du rédacteur (cf. *tag-signature*).

La détection des mots-clés se fait en dehors des éléments <def>, <biblio>, <geoling>, <b> et <i> (grâce à la construction d'une chaîne virtuelle).

Si aucun de ces mots-clés n'est trouvé, le paragraphe est balisé <doc> ; en cas de détection d'un seul ou de plusieurs de ces mots-clés, il est balisé <com> (la reconnais-

sance des paragraphes mixtes ne peut en effet avoir lieu à ce moment de l'algorithme, cf. infra).

Si l'article contient un seul paragraphe, celui-ci est balisé directement <mixt>, sans vérification de son contenu.

Les articles des volumes 21-23 font l'objet d'un traitement particulier : les paragraphes sont tous balisés <doc> (et cela même s'ils sont très souvent mixtes), sauf s'ils ne comprennent aucune étiquette géolinguistique : dans ce cas, ils sont balisés <com>. Ce traitement s'applique y compris si l'article ne contient qu'un seul paragraphe.

5.4.11.4 Validation de la séquence de paragraphes

Erreurs de balisage

La méthode de reconnaissance par mots-clés (→ ci-dessus) n'est pas fiable à 100%. Plusieurs erreurs sont possibles :

1/ Des paragraphes balisés <com> sont en réalité des paragraphes au contenu mixte (documentation et commentaire, à baliser <mixt>). En effet, avec la méthode par mots-clés, ces paragraphes mixtes ne sont pas reconnus : ils sont balisés <com>, puisqu'ils contiennent une partie de commentaire, comme dans l'exemple suivant :

> <p>Ebenso aus frequentativus „ein häufiges tun anzeigend" entlehnt als grammatikerausdruck <geoling>mfr.</geoling>nfr.</geoling> fréquentatif „(verbe) qui exprime l'action comme fréquente, répétée" (seit Meigret, 1550).</p> (FEW 3, 776b, FREQUENTATIO)

2/ Des paragraphes balisés <com> sont en réalité des paragraphes de documentation dans lesquels se trouve malencontreusement un terme technique répertorié dans l'une des bases *few-com-wort-base* ou *few-com-mot-base*. Par exemple, l'application de l'algorithme à l'article MACER (FEW 6/1, 5a-8a) a mené au balisage erroné d'un paragraphe de documentation en <com> à cause de la présence du terme *gebraucht* dans la molécule "informations complémentaires" de l'unité minimale de traitement :

> <p>Fr. emmaigrir v. n. „amaigrir" (seit 12. jh. ; noch Ac 1694, dann kaum noch **gebraucht**[14]), pr. emmaigri, Aix s'enmeigri P, mars. enmeigri A, St-AndréV. e m m a y g r í ALLo 693, bearn. enmagri-s ; Alais emmaïgresí „faire maigrir". Mfr. emmeigrissement „amaigrissement" (16. jh.). — Bearn. entremagrat adj. „mêlé de maigre (du lard)". — Agn. megredi m. „Vendredi Saint" Mir agn.</p>

3/ Des paragraphes balisés <doc> sont en réalité des paragraphes <com> qui ne contiennent aucun des termes techniques répertoriés, soit pour cause de non-exhaustivité de la liste de mots-clés utilisée, soit, plus fréquemment, parce qu'il s'agit du dernier paragraphe du commentaire contenant uniquement des références bibliographiques, comme dans l'exemple suivant (FEW 16, 175b, HASAL) :

> <p>Diez 610 ; ML 4067 ; Stephan 62 ; Gam Germ 1, 213.</p>

Possibilités de résolution

Présence de parenthèses. Nous pouvons résoudre certains de ces problèmes en rendant l'algorithme de balisage plus complexe. Une solution très fiable pour résoudre l'erreur numéro 2 (paragraphe de documentation erronément balisé <com>) consiste à utiliser

la présence de parenthèses comme filtre : un mot-clé ne sera pas pris en compte s'il se trouve à l'intérieur de parenthèses.

Fréquence des étiquettes géolinguistiques. Un indicateur qui peut être utile pour reconnaître un champ documentaire est la fréquence des étiquettes géolinguistiques galloromanes. En effet, la documentation contient obligatoirement une de ces étiquettes au moins et, en général, elle en contient un grand nombre. Le commentaire peut également en contenir, mais en nombre généralement moins élevé. Une fréquence élevée de <geoling> (étiquettes géolinguistiques galloromanes) est un indice quasi certain qu'il s'agit d'un paragraphe de documentation. L'utilisation de ce critère de fréquence nécessite le balisage préalable des étiquettes géolinguistiques, ainsi que la distinction entre les étiquettes galloromanes et non galloromanes (distinction effectuée par un balisage différent : les étiquettes galloromanes sont balisées <geoling>, tandis que les autres sont balisées <lang>, cf. *tag-geoling*). Le seuil de fréquence choisi pour que le nombre de <geoling> soit considéré comme élevé a été calculé à la suite de tests réalisés sur le corpus. Les données de chaque paragraphe du corpus de test (nombre de mots et nombre de <geoling>) ont été examinées et ont fait apparaître un seuil permettant dans beaucoup de cas de distinguer paragraphes de documentation et paragraphes de commentaire[15]. Ce critère statistique peut donc servir comme indicateur. Il n'est toutefois pas totalement fiable, surtout en cas de paragraphe très court. En outre, il reste coûteux algorithmiquement s'il doit être calculé pour chaque paragraphe.

Séquençage des paragraphes. Après application de l'algorithme décrit ci-dessus, nous disposons d'une information intéressante : tous les paragraphes de l'article ayant été balisés <doc>, <com> ou <mixt>, la configuration générale de l'article se dessine. Il est donc possible de se servir du séquençage des paragraphes pour détecter des erreurs de balisage.

En effet, le cas généralement attendu est une suite de paragraphes <doc> suivie d'une suite de paragraphes <com>. Certes, tous les articles ne respectent pas ce schéma : le commentaire peut soit être absent, soit être inclus à l'intérieur de paragraphes de documentation (paragraphes mixtes). Ces cas particuliers (et fréquents uniquement dans les volumes 1, 2/1 et 3) mis à part, si la séquence de paragraphes ne suit pas le schéma "doc-com", il est fort probable que ce soit en raison d'erreurs dues aux limitations des possibilités de détection de la documentation et du commentaire.

Les seuls cas demandant inspection sont donc les articles contenant plus d'un paragraphe et ne respectant pas le schéma doc-com. Quatre situations sont suspectes : (1) la présence d'un paragraphe de documentation en fin d'article ; (2) la présence de paragraphes de commentaire au milieu de paragraphes de documentation ; (3) l'absence totale de paragraphes de documentation (indice de paragraphes mixtes) ; (4) l'absence totale de paragraphe de commentaire (situation en rapport avec la première). Le calcul du seuil de fréquence des étiquettes géolinguistiques sera donc effectué uniquement pour ces cas suspects.

Remarque. Ces vérifications ne concernent pas les articles des volumes 21 à 23, qui suivent un schéma particulier ($\rightarrow$ 3.5.2).

[15]Il serait sûrement utile, lorsque davantage d'articles seront disponibles, de recommencer cette analyse sur la base d'un grand nombre d'articles pour confirmer le seuil de décision.

Algorithme de validation

Les articles des volumes 21 à 23 ne sont pas concernés par cette partie.

La vérification du balisage s'effectue en trois phases successives.

1. Si le dernier paragraphe est marqué par la balise <doc>, l'algorithme calcule la fréquence des étiquettes géolinguistiques de ce paragraphe. Si la fréquence est en-dessous du seuil requis, le paragraphe est rebalisé <com>.

2. L'algorithme cherche ensuite le dernier paragraphe marqué <doc> de l'article. S'il n'en trouve pas, cela signifie que tous les paragraphes de l'article ont été balisés <com>. Dans ce cas, ils sont rebalisés <mixt>, excepté le dernier : au cas où c'était un ancien <doc> rebalisé <com> (cf. phase 1), il reste <com>, sinon, il fait l'objet d'un test supplémentaire de fréquence des étiquettes géolinguistiques. Si celles-ci sont très fréquentes, le paragraphe est rebalisé <mixt> ; dans le cas contraire, il reste marqué <com>. Quel que soit le traitement du dernier paragraphe (donc même s'il existe un commentaire final), l'article est typé "mixt-only" et la phase suivante n'est pas effectuée.

3. Si le dernier paragraphe <doc> de l'article est précédé d'un ou de plusieurs paragraphe(s) marqué(s) <com>, celui-ci ou ceux-ci est/sont directement rebalisé(s) <mixt>.

À la fin de l'algorithme de validation, l'article est typé "doc-com", "doc-mixt" ou "mixt-only" en fonction du balisage final des paragraphes.

5.4.11.5 Résumé

Dépendances

L'algorithme *split-doc-com* dépend du balisage préalable du champ des notes (FFML), des signatures (*tag-signature*), des étiquettes géolinguistiques (*tag-geoling*), des étymons (*tag-etymon*), des définitions (*tag-def*) et de l'entrée (*tag-entry*). Il dépend également du balisage des articles de renvois (*tag-renvoi*), puisqu'il n'est pas appliqué à ces derniers.

Split-doc-com utilise en outre deux listes de mots-clés : *few-com-mot-base* (pour les articles dont la métalangue est le français) et *few-com-wort-base* (pour les articles dont la métalangue est l'allemand).

Description

L'algorithme *split-doc-com* n'est pas appliqué aux articles de renvoi. Dans les autres articles, il balise chaque paragraphe (hors entrée, et excepté les paragraphes de note) en <doc> (paragraphe de documentation), <com> (paragraphe de commentaire) ou <mixt> (paragraphe contenant à la fois de la documentation et du commentaire). L'identification des paragraphes de commentaire est basée sur deux listes de mots-clés, l'une pour les articles dont la métalangue est le français (*few-com-mot-base*), l'autre pour les articles dont la métalangue est l'allemand (*few-com-wort-base*).

Les articles ne contenant aucun paragraphe sont typés "entry-only". Ceux contenant un seul paragraphe sont typés "mixt-only", et le paragraphe en question est balisé <mixt>. Tous les articles des volumes 21 à 23 sont typés "doc-concept" et reçoivent un traitement particulier : les paragraphes sont balisés <doc>, sauf s'ils ne contiennent aucune étiquette géolinguistique <geoling> : dans ce cas, ils sont balisés <com>.

Les articles contenant plus d'un paragraphe (hors volumes 21-23) sont traités en deux phases successives. La première phase consiste en l'application de l'algorithme de balisage expliqué ci-dessus. La seconde phase, consistant en un algorithme de validation, vérifie ensuite ce balisage (en trois sous-phases, cf. ci-dessus). À la fin de l'algorithme de validation, un attribut *type* est ajouté à la balise <art> en fonction du balisage final. Si la séquence de paragraphes se compose d'au moins un paragraphe <doc> suivi d'au moins un paragraphe <com>, l'article est typé "doc-com". Si l'article contient uniquement des paragraphes <doc>, il est typé "doc-only". Sinon, il est typé "doc-mixt".

5.4.12 Balisage des marqueurs alphanumériques : tag-numbering

<p>3. <geoling>Mfr.</geoling> <i>pluraliste</i> adj. <def>„variable, changeant (de caractère)“</def> (hap., <biblio>Lac</biblio>).	<p><pnum id="I 3">3.</pnum> <geoling>Mfr.</geoling> <i>pluraliste</i> adj. <def>„variable, changeant (de caractère)“</def> (hap., <biblio>Lac</biblio>).

(FEW 9, 101a, PLURALIS)

5.4.12.1 Objectifs de l'algorithme

L'algorithme *tag-numbering* a deux objectifs. Le premier consiste à reconnaître les marqueurs alphanumériques présents au début des paragraphes de documentation d'un article du FEW. Ces marqueurs sont balisés au moyen de l'élément XML <pnum>. Un attribut *id* rétablit la totalité de la numérotation, tant explicite qu'implicite, de la façon suivante :

> <p><pnum id="I 3">3.</pnum>

Le second objectif de *tag-numbering* est de reconnaître, dans les paragraphes de commentaire de l'article, les marqueurs alphanumériques qui renvoient aux marqueurs de la documentation, par exemple entre parenthèses dans l'extrait suivant (FEW 9, 101a, PLURALIS) :

> Aus der ersten bed. wurde es schon sehr früh entlehnt (I 1), zuerst unter angleichung des suffixes an dessen volkssprachliche form (a).

Ces marqueurs sont balisés au moyen de l'élément XML <pref> :

> Aus der ersten bed. wurde es schon sehr früh entlehnt (<pref id="I 1">I 1</pref>), zuerst unter angleichung des suffixes an dessen volkssprachliche form (<pref id="I 1 a">a</pref>)

Outre les raisons évoquées pour la modélisation du FEW (→ 3.7.4), le balisage

des marqueurs alphanumériques est nécessaire à l'application d'autres algorithmes de rétroconversion. Il est notamment utilisé pour le balisage de la microstructure (cf. *tag-microstructure*).

5.4.12.2 Indicateurs

Marques de numérotation

> Les marqueurs alphanumériques ont été examinés par Büchi (1996, 100-104). Ils sont constitués au maximum de six niveaux hiérarchisés : chiffres romains, chiffres arabes, lettres romaines, lettres grecques, lettres romaines prime, lettres grecques prime.

Ces *marques de numérotation* constituant des ensembles finis, il est possible de les répertorier dans des listes. Nous constituons six listes contenant chacune toutes les marques de numérotation appartenant à un même niveau. Dans chacune de ces listes, les marques sont ordonnées de la plus haute à la plus basse, ce qui donne la hiérarchie suivante :

1. few-numbering-roman-numbers-base (I, II,..., X)
2. few-numbering-arabic-numbers-base (0, 1,..., 99)
3. few-numbering-roman-letters-base (a, b,..., z)
4. few-numbering-greek-letters-base (alpha, beta, ..., pi)
5. few-numbering-roman-letters-prime-base (a', b',..., z')
6. few-numbering-greek-letters-prime-base (alpha', beta', ..., pi')

L'ensemble des séquences possibles de marques hiérarchisées constitue ce que nous appellerons la *hiérarchie canonique de numérotation du FEW*. Par exemple, I.1.b. ou I.a. appartiennent à la hiérarchie canonique de numérotation du FEW, tandis que 1.II. n'en fait pas partie.

Remarque. La liste *few-numbering-roman-numbers-base* contient des cas de collisions avec les listes *few-etymon-base* (2 collisions : « I », « X ») et *few-bib-base* (1 collision : « V »). La liste *few-numbering-roman-letters-base* contient en outre des cas d'enchâssements avec la liste *few-geoling-base* (« d. », « e. »). Ces cas problématiques devront être résolus.

Repères de numérotation

> Dans la documentation, certains paragraphes débutent par une séquence de marques, qui permettent d'identifier ce paragraphe (et éventuellement les suivants) de façon unique au sein de la structure de l'article. Dans une séquence, les marques sont séparées par un point (cf. figure 5.8).

I. 1. a. Fr. *personne* f. „individu, homme ou femme“ [1]) (seit ca. 1180), *piersonne* (pik. 13. jh.,

FIGURE 5.8 – Exemple de repère de numérotation (FEW 8, 268b, PĔRSŌNA)

wörter, die dinge oder wesen bedeuten. Daraus im 14. jh. entlehnt 2 a. Die verwendung des wortes, wie sie sich bei Priscian findet, wird im 16. jh. dazu entlehnt (b). In der ursprünglichen bed. „auf sich selbst beruhend“ wird es im 19. jh. von der technik entlehnt (3).

FIGURE 5.9 – Exemple de références de numérotation (FEW 12, 357a, SUBSTANTIVUS)

Nous appellerons dorénavant *repère de numérotation* une séquence de marques de numérotation située au début d'un paragraphe de documentation. Un repère de numérotation se définit comme une suite de marques séparées par un point (et une espace éventuelle), telle que chacune des marques appartient à une liste de niveau inférieur à celle de la marque précédente.

L'ensemble des repères de numérotation apparaissant dans la documentation d'un article constitue un sous-ensemble de la hiérarchie canonique de numérotation du FEW. Nous appellerons cet ensemble la *hiérarchie effective de numérotation d'un article*. Par exemple, si un article contient deux niveaux de hiérarchisation marqués par les chiffres arabes 1 et 2 et par les lettres romaines a et b, ces lettres n'apparaissant qu'à l'intérieur du niveau 1, la hiérarchie effective de l'article comprend les trois séquences « 1.a. », « 1.b. » et « 2. ».

Notons que la hiérarchie effective d'un article est indépendante de la façon dont se présentent les repères dans la structure de surface du FEW. La règle d'économie qui est de mise dans le FEW cause notamment l'effacement en structure de surface de certains marqueurs, qui deviennent donc implicites : dans notre exemple, le repère « 1.b. » apparaîtrait dans le FEW (jusqu'au changement de rédaction à dater du volume 25) sous la forme d'un « b. » uniquement, sans répétition du « 1. » (un exemple réel de ce cas se trouve dans FEW 16, 130a, HALM 1 b).

Références de numérotation

Dans le commentaire apparaissent des marques de numérotation qui reprennent certains repères de la documentation et qui y réfèrent. Ces *références de numérotation* sont similaires aux repères de numérotation, à trois différences près.

> Tout d'abord, les marques composant une référence ne sont en général pas séparées par un point. Ensuite, les références de numérotation se trouvent dans les paragraphes de commentaire (ou de notes) uniquement (et non dans les paragraphes de documentation). Enfin, les références de numérotation peuvent apparaître n'importe où dans le commentaire et dans n'importe quel ordre, sans suivre de règles systématiques. Néanmoins, elles apparaissent souvent entre parenthèses et/ou en début ou en fin de phrase (cf. figure 5.9).

Les références de numérotation référant aux repères de numérotation, elles ne peuvent contenir que les marques qui sont effectivement présentes dans ces derniers. Leur reconnaissance nécessite donc le balisage préalable des repères de numérotation et la constitution de la hiérarchie effective de l'article, afin de connaître la *liste effective des marques de numérotation* autorisées dans l'article traité. Cette liste de marques contient toutes les marques de numérotation apparaissant dans les repères de numérotation de l'article (dans notre exemple, la liste de marques contiendra les quatre marques 1, 2, a et b).

> Les marques de numérotation apparaissant dans le commentaire peuvent être ambiguës. La marque « a », par exemple, pourrait être une forme du verbe avoir dans les commentaires en français. Il y a également ambiguïté possible avec les marqueurs associés à un étymon de renvoi.

Les ambiguïtés avec les marqueurs associés à un étymon de renvoi sont résolus par l'application de *tag-renvoi* avant *tag-numbering* dans la séquence de rétroconversion. Les ambiguïtés avec des mots du français ou de l'allemand sont réduites en partie par l'examen du contexte (parenthèses et ponctuation).

5.4.12.3 Reconnaissance des marqueurs dans la documentation

La reconnaissance des repères de numérotation se base sur les indicateurs positionnels (position en début de paragraphe) et textuels (recherche de marques appartenant aux six listes de numérotation). Une suite de caractères est un repère de numérotation si elle se trouve en début d'un paragraphe <doc> ou <mixt> et si elle est composée de marques de numérotation appartenant aux six listes de numérotation, ces marques étant séparées par des points ou des espaces uniquement. La séquence doit en outre appartenir à la hiérarchie canonique de numérotation du FEW.

L'algorithme consiste donc, dans chaque paragraphe <doc> ou <mixt>, à rechercher en début de paragraphe les marques présentes dans les listes, en veillant à ce qu'elles soient séparées uniquement par des points ou des espaces. Si une séquence de marques valide (c'est-à-dire appartenant à la hiérarchie canonique du FEW) est détectée, celle-ci est balisée et ajoutée à la hiérarchie effective de numérotation de l'article.

Il est important de signaler que la recherche des séquences se fait obligatoirement dans un chaîne virtuelle, qui rend transparentes la plupart des balises et notamment les balises <geoling>. Ce procédé permet de résoudre les cas de collisions entre les listes de numérotation et la liste *few-geoling-base*. En effet, l'algorithme de détection des repères de numérotation est assez fiable pour désambiguïser des éléments comme « d », « e » etc. qui sont considérés comme des sigles géolinguistiques par *tag-geoling* (faux positifs) à cause de leur appartenance à la liste de mots-clés *few-geoling-base*. Après balisage des repères de numérotation, les éléments <geoling> apparaissant à

l'intérieur de ceux-ci sont simplement supprimés. L'extrait suivant, balisé

> <p><geoling>d.</geoling> <geoling>Mfr.</geoling> <geoling>nfr.</geoling> <i>pluriel</i> m. <def>„nombre pluriel"</def>

avant l'application de *tag-numbering*, devient donc, après application de ce dernier :

> <p><pnum id="I 1 d">d.</pnum> <geoling>Mfr.</geoling> <geoling>nfr.</geoling> <i>pluriel</i> m. <def>„nombre pluriel"</def> (FEW 9, 101a, PLURALIS)

5.4.12.4 Reconnaissance des marqueurs dans le commentaire

La reconnaissance des références de numérotation présentes dans le commentaire s'appuie sur la détection des repères de numérotation réalisée précédemment et, plus précisément, sur la liste des marques de numérotation contenues dans ces repères (liste effective des marques de numérotation). Ces marques de numérotation sont recherchées dans le commentaire de l'article. Leur contexte est ensuite examiné afin de les filtrer :

1. Une marque de numérotation doit obligatoirement, soit ouvrir le paragraphe, soit être directement précédée d'une espace, d'un point ou d'une parenthèse ouvrante. Si ce n'est pas le cas, elle n'est pas considérée comme une marque de numérotation.

2. Une marque de numérotation doit être directement suivie d'une espace, d'un point, d'un point-virgule, d'une virgule ou d'une parenthèse fermante. Si ce n'est pas le cas, elle n'est pas considérée comme marque de numérotation.

Si la marque est un nombre, deux vérifications supplémentaires sont effectuées :

1. Si elle est directement précédée de la séquence espace + « n » (pour *note*)+ espace, elle n'est pas considérée comme une marque de numérotation ;

2. Si elle est suivie d'une virgule elle-même directement suivie d'une espace facultative et d'un nombre (l'expression régulière correspondante est ", ?[0-9]+"), elle n'est pas considérée comme une marque de numérotation.

Les marques qui passent avec succès ces examens contextuels sont considérées comme des marques de numérotation. Elles sont alors examinées à nouveau en prenant en compte leur ordre d'apparition, afin de rétablir l'implicite. Par exemple, si la marque « a » apparaît dans le texte fewien et que la marque « 1 » apparaissait précédemment, l'algorithme vérifie que la séquence « 1.a. » appartient aux repères de numérotation de la liste des repères de l'article. Le cas échéant, la marque « a » est balisée <pref id="1 a">a</pref>. Les marques qui ne peuvent être décodées sont balisées également, mais un attribut *status*="ambigu" est ajouté à la balise <pref> pour signaler une erreur possible et permettre une vérification manuelle.

5.4.12.5 Résumé

Dépendances

La reconnaissance des marqueurs alphanumériques nécessite le balisage de la documentation et du commentaire (cf. *split-doc-com*).

Tag-numbering utilise en outre six listes de mots-clés (few-numbering-*-base) reprenant tous les marqueurs alphanumériques.

Les collisions et enchâssements de *few-geoling-base* et *few-bib-base* avec les listes few-numbering-*-base sont résolues par la création d'une chaîne virtuelle qui, pendant la recherche des marqueurs, rend transparentes les balises <geoling>, <lang> et <biblio>, de sorte que ces dernières peuvent ensuite être détectées et supprimées après le balisage des marqueurs.

Description

L'algorithme *tag-numbering* a pour objectifs (1) de baliser (<pnum>) les repères de numérotation présents dans la documentation d'un article du FEW et (2) de reconnaître et de baliser (<pref>) les références de numérotation présentes dans le commentaire d'un article du FEW. Le traitement d'un article s'effectue en deux phases successives, correspondant à ces deux objectifs. Les articles de renvoi ne sont pas traités.

L'algorithme utilise des structures de données adéquates permettant notamment de mémoriser la liste des repères de numérotation détectés dans la documentation (hiérarchie effective de numérotation de l'article) et la liste des marques de numérotation appartenant à ces repères (liste effective des marques de numérotation de l'article), afin de les utiliser comme indicateurs lors de la détection des références de numérotation présents dans le commentaire.

1. Détection et balisage des repères de numérotation en <pnum>

Chaque paragraphe <doc> et chaque paragraphe <mixt> est examiné après construction d'une chaîne virtuelle dans laquelle la plupart des balises sont rendues transparentes. Le traitement s'effectue en deux étapes : d'abord, la détection et le balisage (<pnum>) du repère de numérotation éventuellement présent en début de paragraphe ; ensuite, la suppression systématique des balises <geoling> qui se trouveraient dans ou immédiatement autour des balises <pnum> insérées.

La détection des repères de numérotation s'effectue en itérant sur la séquence de marques de numérotation et de séparateurs (points et espaces) trouvée au début de la chaîne virtuelle. Tous les mots-clés des six listes de numérotation sont recherchés, en privilégiant les marques les plus longues (a prime, par exemple, l'emporte sur a). La séquence de marques détectées est insérée en tant que nouveau repère de numérotation dans la hiérarchie effective de l'article. Le balisage de la séquence s'effectue en intégrant le point qui la suivrait éventuellement.

La seconde étape consiste à supprimer de façon systématique toute balise <geoling>, <lang> ou <biblio> (ainsi que leurs correspondants fermants) trouvée à l'intérieur (ou directement autour) de l'élément <pnum> balisé lors de la première étape.

2. Détection et balisage des références de numérotation en <pref>

Chaque paragraphe marqué <com> est traité après construction d'une chaîne virtuelle rendant transparentes la plupart des balises (dont <geoling>, <lang> et <biblio>) et invisibles les balises <appelnote>, <e> et <def>. Le traitement a lieu en quatre étapes : (1) détection dans le paragraphe de toutes les marques de numérotation préalablement mémorisées dans la liste de marques de l'article, (2) regroupement des marques détectées en séquences de marques, (3) identification des références de numérotation valides parmi les séquences créées, (4) suppression systématique des balises <geoling> qui se trouveraient à l'intérieur (ou directement autour) des éléments <pref> balisés.

La première étape consiste à détecter dans le paragraphe toutes les marques appartenant à la liste effective des marques de numérotation (liste contenant toutes les marques détectées lors du balisage des paragraphes de documentation).

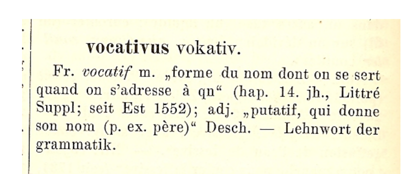
La deuxième étape consiste à analyser les marques en fonction de leur contexte, afin d'éliminer celles qui ne sont pas des marques. L'algorithme itère sur les marques, vérifie leur contexte et détecte les séquences de marques consécutives qui pourraient être des références de numérotation.

La troisième étape consiste à identifier, parmi les séquences conservées, celles qui sont effectivement des références de numérotation, en résolvant l'implicite éventuel. Le décodage de l'implicite se fonde sur les références de numérotation précédemment identifiées dans le paragraphe. Si la séquence décodée appartient à la hiérarchie effective de numérotation de l'article (constituée lors du balisage des repères de numérotation), elle est balisée <pref>. Si la séquence décodée n'appartient pas à la hiérarchie effective de numérotation, ou si la séquence ne peut être décodée, l'algorithme vérifie si la séquence non décodée appartient à la hiérarchie effective de numérotation : le cas échéant, la séquence est balisée <pref>. Si la séquence non décodée n'appartient pas à la hiérarchie effective de numérotation, elle est balisée <pref> avec un attribut *status* indiquant une ambiguïté.

La quatrième étape consiste à supprimer les balises <geoling>, <lang> et <biblio> qui se trouveraient à l'intérieur (ou autour) d'un élément <pref>. Puisque l'algorithme de balisage des références de numérotation est susceptible d'erreurs, un avertissement est émis pour vérification par un linguiste.

5.4.13 Séparation de la documentation et du commentaire (2) : split-mixt-art

<table>
<tr>
<td><mixt>
<p>Fr. <i>appellatif</i> adj. „(nom) qui convient à toute une espèce et non à un individu seul“ (ca. 1350–Ac 1878, GlC ; ‚on dit plutôt <i>nom commun'</i> Ac 1932), mfr. <i>appellatif</i> m. „nom commun“ (ca. 1550). — Lehnwort.</p>
</mixt></td>
<td><mixt mixt-type="mixt_doc">
<p>Fr. <i>appellatif</i> adj. „(nom) qui convient à toute une espèce et non à un individu seul“ (ca. 1350–Ac 1878, GlC ; ‚on dit plutôt <i>nom commun'</i> Ac 1932, mfr. <i>appellatif</i> m. „nom commun“ (ca. 1550). </p>
</mixt>
<mixt mixt-type="mixt_com">
<p>— Lehnwort.</p>
</mixt></td>
</tr>
</table>



vocativus vokativ.
Fr. *vocatif* m. „forme du nom dont on se sert quand on s'adresse à qn“ (hap. 14. jh., Littré Suppl; seit Est 1552); adj. „putatif, qui donne son nom (p. ex. père)“ Desch. — Lehnwort der grammatik.

FIGURE 5.10 – Exemple de paragraphe mixte où les matériaux et le commentaire sont séparés par un tiret (FEW 14, 588b, VOCATIVUS)

(FEW 25, 32b, APPELLATIVUS)

5.4.13.1 Objectifs de l'algorithme

L'algorithme *split-mixt-art* a pour objectif de reconnaître, dans les paragraphes mixtes d'un article, la partie documentaire et la partie consacrée au commentaire. Ces deux parties doivent être séparées par *split-mixt-art* en deux paragraphes distincts, à condition que le texte du FEW rende cette division possible.

Le balisage <mixt> présent par défaut est remplacé par un balisage <mixt mixt-type="mixt-doc"> (au début du paragraphe de documentation) et <mixt mixt-type="mixt-com"> (au début du paragraphe de commentaire) ainsi que par leurs correspondants fermants.

Outre le fait qu'elle relève d'un des paramètres structurels les plus fondamentaux de la microstructure du FEW, la séparation des champs de la documentation et du commentaire au sein d'un paragraphe mixte présente un intérêt indirect de taille. En effet, la détection de certaines informations et, notamment, des unités minimales de traitement, se fait uniquement dans la documentation (cf. *tag-unit*).

5.4.13.2 Indicateurs de reconnaissance

Les paragraphes mixtes ont été balisés <mixt> par l'algorithme *split-doc-com* dans deux cas de figure : soit parce qu'ils constituaient l'unique paragraphe d'un article (susceptible dès lors de contenir à la fois les matériaux et un commentaire intégré), soit parce qu'ils contenaient un terme technique spécifique au champ du commentaire, mais que leur position dans la séquence de paragraphes de l'article était suspecte (cf. *split-doc-com*).

La question est à présent de déterminer comment détecter dans ces paragraphes une éventuelle séparation nette entre une partie documentaire et une partie de commentaire, pour autant que ces deux parties y soient présentes.

> ITy. En cas de paragraphe mixte présentant d'abord la documentation et se terminant par un commentaire, ce dernier est généralement séparé des matériaux par un tiret cadratin ou semi-cadratin (cf. figure 5.10).

derivare ableiten.

Daraus entlehnt fr. *dériver* „faire sortir (une eau) de son lit; faire sortir un mot d'un autre" (seit 12. jh.). — Ebenso entlehnt aus DERIVATIO fr. *dérivation* „action de dériver" (seit 1377); aus DERIVATIVUS mfr. nfr. *dérivatif* „qui dérive qch (t. de méd.)" (seit 1527).

FIGURE 5.11 – Exemple de paragraphe mixte où le tiret sépare deux ensembles constitués chacun d'unités lexicales et d'un bref commentaire (FEW 3, 49b, DERIVARE)

> Toutefois, un tiret (semi-)cadratin en fin de paragraphe peut servir également à introduire un ensemble d'unités lexicales (cf. 13/1, 302b, THĒCA), des références bibliographiques (cf. FEW 16, 172b, HARMSKARA) ou la signature du ou des rédacteur(s) (cf. FEW 25, 882a, AUGMENTATOR). De plus, le tiret peut aussi séparer deux ensembles constitués de matériaux et d'un (embryon de) commentaire (cf. figure 5.11).

Ainsi la présence d'un tiret (semi-)cadratin ne constitue pas, prise isolément, un critère de reconnaissance fiable pour séparer la partie dédiée au commentaire de la partie documentaire. Ce critère doit être couplé à la détection d'un terme technique spécifique au commentaire.

L'algorithme consiste dès lors à détecter dans le paragraphe la présence de mots-clés de la liste *few-com-wort-base* ou *few-com-mot-base* (en fonction de la métalangue de l'article), exactement de la même façon que dans l'algorithme *split-doc-com*, et ensuite à vérifier la présence d'un tiret (semi-)cadratin à gauche (immédiatement ou non, la limite à gauche étant le début du paragraphe) de ce mot-clé.

Deux remarques très importantes sont à prendre en compte :

1. la détection des tirets doit, tout comme la détection des mots-clés, se faire hors des parenthèses, afin d'éviter de retenir des tirets situés dans des fourchettes de datation ;

2. en cas de détection de plusieurs tirets à gauche du mot-clé, c'est le tiret le plus à droite (donc le plus proche du mot-clé) qui doit être retenu (cf. FEW 16, 176a, HASCHEN).

Afin de remédier à des erreurs dues au FEW ou dues à la saisie du texte fewien, les tirets courts sont également pris en compte.

Le tiret le plus proche à gauche du mot-clé et non situé entre parenthèses constitue le délimiteur entre la partie documentation (du début du paragraphe jusqu'au tiret

exclus) et la partie commentaire (du tiret inclus jusqu'à la fin du paragraphe).

> Dans certains paragraphes mixtes, les deux parties ne sont pas séparées du tout : la documentation est intégrée dans des phrases de commentaire. Ces paragraphes peuvent contenir des tirets (semi-)cadratins qui n'introduisent pas un commentaire, mais séparent deux phrases au contenu mixte (exemple FEW 3, 49b, DERIVARE, cf. ci-dessus). Ces cas concernent uniquement les volumes 1, 2/1 et 3 du FEW.

L'algorithme décrit ci-dessus, qui utilise uniquement la détection de mots-clés et la présence de tirets, balisera erronément ces paragraphes mixtes : la partie contenant un mot-clé et située à droite du tiret sera interprétée comme un commentaire, alors qu'elle contient également de la documentation.

Afin d'éviter ces erreurs de balisage, nous ajoutons à l'algorithme un critère de détection supplémentaire : la présence d'étiquettes géolinguistiques galloromanes. Une partie d'un paragraphe mixte pourra être considérée comme un commentaire uniquement si elle ne contient pas d'élément <geoling> (toujours hors des parenthèses, tout comme la détection des tirets et des mots-clés). Cette règle rend la détection d'une partie commentaire au sein d'un paragraphe mixte plus restrictive que dans la détection de paragraphes de commentaire effectuée par *split-doc-com*, où le commentaire pouvait contenir un nombre bas d'étiquettes géolinguistiques (cf. calcul de fréquence des <geoling> dans *split-doc-com*). Le calcul de fréquence des étiquettes géolinguistiques n'est pas applicable dans les paragraphes mixtes, car ces derniers sont très courts ; or, la fiabilité du calcul de fréquence décroît avec la longueur du texte pris en compte.

5.4.13.3 Balisage

Pour faciliter la mise au point des algorithmes qui suivent *split-mixt-art* dans la séquence de rétroconversion, les deux parties sont séparées en deux paragraphes distincts, tous deux balisés <mixt> avec un attribut *mixt-type* indiquant s'il s'agit d'une partie de documentation ou de commentaire. La partie documentaire est balisée <mixt mixt-type="mixt_doc"> et la partie commentaire <mixt mixt-type="mixt_com">.

Les deux parties seront regroupées en un seul paragraphe par un algorithme exécuté en fin de séquence (cf. *merge-mixt-p*).

Par ailleurs, un avertissement est émis à la fin de l'algorithme pour tous les paragraphes <mixt> qui n'ont pas été séparés par *split-mixt-art*, afin de permettre une vérification par un expert humain et, éventuellement, un balisage manuel suivi d'une reprise de la séquence de rétroconversion à partir du fichier XML requis.

5.4.13.4 Résumé

Dépendances

L'algorithme *split-mixt-art* dépend du balisage préalable des paragraphes mixtes (*split-doc-com*) et des étiquettes géolinguistiques (*tag-geoling*).

Split-mixt-art utilise en outre deux listes de mots-clés : *few-com-mot-base* (pour les articles dont la métalangue est le français) et *few-com-wort-base* (pour les articles dont

la métalangue est l'allemand).

Description

L'algorithme *split-mixt-art* s'applique uniquement aux articles contenant des paragraphes mixtes, c'est-à-dire des articles dont le type (attribut délivré par *split-doc-com*) est "mixt-only" ou "doc-mixt". Dans ces articles, il repère les paragraphes mixtes (balisés <mixt>) et tente de les séparer en deux sous-paragraphes : le premier (typé "mixt_doc") contient une partie identifiée comme documentation et le second (typé "mixt_com") contient une partie identifiée comme commentaire.

Le traitement d'un paragraphe est effectué sur une chaîne virtuelle dans laquelle la plupart des balises sont transparentes et certaines invisibles (cf. partition). En outre, le texte situé entre des parenthèses n'est pas pris en compte.

L'algorithme se déroule en trois phases. Il cherche d'abord les mots-clés de *few-com-mot-base* ou de *few-com-wort-base* (en fonction de la valeur de l'attribut *lang* de <art>). Il détermine ensuite où se trouve exactement la séparation entre les deux parties documentation et commentaire, en cherchant un délimiteur. Enfin, il vérifie qu'il n'y a aucune balise <geoling> entre le délimiteur et la fin du paragraphe. Si l'une des trois phases n'est pas concluante, le paragraphe n'est pas subdivisé.

La séparation entre les deux parties est déterminée en examinant le contexte gauche du mot-clé (le mot-clé pris en compte étant celui situé le plus à gauche dans le texte) et en y détectant des tirets. Le tiret situé le plus à droite (donc le plus près du mot-clé) est considéré comme délimiteur entre les deux parties.

5.4.14 Balisage des affixes : tag-affix

<p><note id="2">2) Ist diese form als vermeintlicher suffw. (*-<sc>ŏlu</sc> für -<sc>ālis</sc>) zu erklären ?	<p><note id="2">2) Ist diese form als vermeintlicher suffw. (*<affix type="suffix">-<sc>ŏlu</sc></affix> für <affix type="suffix">-<sc>ālis</sc></affix>) zu erklären ?

(FEW 6/1, 53a, MAIĀLIS n. 2)

5.4.14.1 Objectifs de l'algorithme

L'algorithme *tag-affix* a pour objectif de baliser les préfixes et suffixes que le FEW identifie comme tels, qu'il s'agisse d'affixes hérités de la protolangue ou d'affixes créés dans les différentes langues de la Galloromania.

Conformément à la modélisation proposée, les affixes doivent être balisés comme suit :

<affix type="prefix">Ici se trouve un préfixe</affix>

5.4.14.2 Indicateurs textuels

> ITe. Les affixes étymologiques (provenant de langues diverses) ont fait l'objet à l'ATILF d'une répertorisation qui a débouché sur une liste comptant 46 préfixes (*ab-*, *ante-*, *cum-*, *trans-* etc.) et 230 suffixes (*-abundu*, *-enko*, *-uculu* etc.).

Le répertoire des affixes a été transformé en deux listes de mots-clés intitulées *few-prefix-base* et *few-suffix-base*. Les tirets marquant la limite entre morphèmes y sont conservés. Les caractères spéciaux (notamment les voyelles brèves et longues) respectent les normes de codage établies pour la saisie du FEW.

Afin de remédier aux incohérences du FEW ou de la saisie, la recherche des mots-clés s'effectue en désactivant les signes diacritiques, par la méthode d'aplatissement des caractères (→ 4.2.2). Les mots-clés sont recherchés sans tenir compte de la typographie (petites capitales et italiques), car celle-ci varie en fonction du mot-clé.

5.4.14.3 Indicateurs typographiques

Tous les affixes ne sont pas répertoriés dans les listes *few-prefix-base* et *few-suffix-base*. Il est difficile, notamment, de constituer une liste exhaustive des mentions de suffixes et de préfixes qui constituent dans le FEW non un morphème étymon, mais un morphème appartenant à un lexème galloroman (par exemple les suffixes français *-iste* ou *-ette*).

> IPo. ITy. Les affixes étymologiques peuvent apparaître à n'importe quelle position dans un article du FEW. Les affixes galloromans apparaissent le plus souvent dans la partie consacrée au commentaire ou dans les notes d'un article, mais des exceptions existent (FEW 25, 770b, AUCA I 5 d). Ils sont parfois précédés ou suivis d'un discours qui les identifie comme affixes :
>
> > En occitan, la finale a été greffée par le suffixe péjoratif *-aud* (α) ou par *-an* < ? (β). À Lant., par le suffixe péjoratif *-ard* (γ). (FEW 24, 592b, ANIMAL)

La présence, dans le contexte proche (la phrase) de l'affixe, de termes spécifiques tels que *préfixe*, *suffixation* etc. constitue un indicateur positionnel pertinent, mais non suffisant, car peu d'affixes bénéficient de ce marquage contextuel explicite.

> ITy. Lorsqu'ils ont la fonction d'étymon, les affixes sont généralement en petites capitales. Dans la refonte, ils sont parfois situés entre crochets droits, lorsqu'ils jouent le rôle de marqueur de structuration. Ils sont parfois précédés ou suivis d'un signe +.
> Les affixes galloromans apparaissent en italiques et sont, en fonction de leur nature, précédés et/ou suivis d'un tiret.

Les mentions de phonèmes ou de graphèmes apparaissent également en italiques et

peuvent être précédées ou suivies d'un tiret :

> Dans ANIMAL comme dans ANIMALIA, le traitement galloroman presque unanime du groupe -N'M- est *-rm-/-lm-* ; les témoins d'un traitement *-m-* ne sont que sporadiques. (FEW 24, 592b, ANĬMAL)

5.4.14.4 Combinaison des indicateurs

La combinaison des indicateurs textuels, typographiques et positionnels permet d'écrire une règle de reconnaissance des affixes. Un extrait de texte du FEW est un affixe s'il remplit une au moins des deux conditions suivantes :

– il appartient aux listes *few-prefix-base* et *few-suffix-base* ;

– il se présente en italiques, il est précédé et/ou suivi d'un tiret et il se trouve dans le commentaire ou les notes.

Afin d'éviter la détection de faux positifs (phonèmes ou graphèmes), les extraits de texte qui répondraient à toutes ces conditions, mais qui se limiteraient à un seul caractère (outre les tirets), ne sont pas considérés comme des affixes.

5.4.14.5 Résumé

Dépendances

La reconnaissance des affixes nécessite le balisage préalable des sigles géolinguistiques (*tag-geoling*), des marqueurs alphanumériques (*tag-numbering*), des renvois (*tag-renvoi*) et des définitions (*tag-def*).

Tag-affix utilise deux listes de mots-clés : *few-prefix-base* et *few-suffix-base*.

Description

Tag-affix a pour objectif de reconnaître et baliser dans un article du FEW les mots-clés des listes *few-prefix-base* et *few-suffix-base*, ainsi que les affixes non répertoriés qui apparaissent dans les parties de commentaire et de notes.

Les mots-clés des listes *few-prefix-base* et *few-suffix-base* sont détectés dans tous les paragraphes de l'article, y compris les paragraphes de notes. Le champ de l'entrée est exclu du traitement. Chaque paragraphe est traité deux fois : une première fois avec la liste *few-prefix-base*, une seconde fois avec la liste *few-suffix-base*. La détection des mots-clés se fait à chaque fois dans une chaîne virtuelle dans laquelle la plupart des balises sont transparentes et plusieurs invisibles avec leur contenu. L'invisibilité de <affix> est entre autres requise.

Les préfixes immédiatement suivis d'une fin de ligne (<lb\>) ne sont pas balisés.

Les affixes non répertoriés sont détectés uniquement dans les parties de commentaire (<com> et <mixt>) et les notes. Le traitement se fait dans une chaîne virtuelle rendant transparentes ou invisibles la plupart des balises (cf. partition). Les éléments en italiques qui sont visibles dans la chaîne virtuelle ainsi créée sont passés en revue.

Lorsqu'ils contiennent plus d'un caractère ainsi qu'un tiret en début et/ou fin d'élément, ils sont balisés <affix>.

Les affixes balisés reçoivent un attribut *type* dont la valeur est "prefix", "suffix" ou " ?" (ce dernier uniquement pour les affixes contenant un tiret à la fois en début et en fin de chaîne).

5.4.15 Balisage des dates : tag-date

<p>1. Afr. mfr. <i>génitif</i> „qui a la faculté d'engendrer" (13<e>e</e> s. —1615), <i>genetif</i> (1464)	<p>1.Afr. mfr.<i>génitif</i> „qui a la faculté d'engendrer" (<date>13<e>e</e> s.</date> —<date>1615</date>), <i>genetif</i> (<date>1464</date>)

(FEW 4, 102b, GENITIVUS 1)

5.4.15.1 Objectifs de l'algorithme

L'algorithme *tag-date* a pour objectif de reconnaître la molécule de l'infrastructure qui concerne les datations des lexèmes. Ces datations sont balisées <date> (→ 3.6.4) :

<date>1956</date> ; <date>15 jh.</date>

Outre son utilité directe pour l'utilisateur, le balisage des datations présente un intérêt indirect dans le processus de rétroconversion, dans la mesure où il participe à la reconnaissance du composant de la cellule lexicale que nous avons nommé "précisions" (→ 3.6.4).

5.4.15.2 Indicateurs de reconnaissance

> IPo. Les datations n'apparaissent pas dans le champ de l'entrée, mais bien dans le reste de l'article (documentation, commentaire et notes). Dans la documentation, les dates apparaissent à l'intérieur des parenthèses de précision.
> ITe. Les datations peuvent apparaître sous des formes variées : date exacte (1956), indication de siècle (12. jh.), fourchette de deux siècles (12.-14. jh.), partie de siècle (ende 14. jh.) etc.

Nous avons défini quatre formats de datation, chacun étant représenté par une expression régulière. Chacun de ces formats se présente sous deux formes, selon que la métalangue de l'article est l'allemand ou le français. Les expressions régulières sont recherchées dans une chaîne virtuelle où les balises <e>. . .</e> sont transparentes.

Date exacte

Le premier format de datation est une date exacte, c'est-à-dire une séquence de 4 chiffres allant de 1000 à 1999. Cette limite à 1999 nous paraît suffisante pour la rétroconversion du FEW imprimé, sachant que le dernier fascicule du volume 25 est paru en 2002[16].

Les dates exactes peuvent être précédées des mentions *ca.* ou *env.*, qui sont intégrées dans le balisage. L'expression régulière est la suivante (le caractère _ représente une espace) :

`/(ca\._?|env\._?)?(1[0-9]{3})/`

Siècle

Le deuxième format est une indication de siècle, c'est-à-dire un nombre entre 9 et 21 suivi de jh. (pour jahrhundert) ou de s. (pour siècle). Les expressions régulières sont :

- pour l'allemand : `/(9|1[0-9]|2[0-1])\._?jh\./`
- pour le français : `/((9)|(1[0-9])|(2[0-1]))e_?(s\.|siècle)/`

Fourchette

Le troisième format est une fourchette de deux siècles (séparés par un tiret ou un tiret (semi-)cadratin). En français, on trouve aussi des mentions de deux siècles coordonnés par et. Les expressions régulières sont les suivantes :

- pour l'allemand :
  `/((9|1[0-9]|2[0-1])\.)_?[----]?((9|1[0-9]|2[0-1])\._?jh\.)/`
- pour le français :
  `/((9|1[0-9]|2[0-1])e)_?([----]|et)_?(((9|1[0-9]|2[0-1])e)_?(s\.|siècle))/`

Partie de siècle

Le quatrième format est une partie de siècle, telle que "1. hälfte 13. jh." (FEW 9, 307a, PRAESENS 3). Les expressions régulières sont les suivantes :

- pour l'allemand :
  `/([1234]_?\._?viertel|[123]_?\._?drittel|[12]_?\._?hälfte|anfang_?|mitte_?|ende _?)((9|1[0-9]|2[0-1])\._?jh\.)/`
- pour le français :
  `/((1er|2e|3e)_?t\._?|(1ère|2e)_?m\._?|(1er|[234]e)_?q\._?|mil\._?|déb\._?|début_?|fin_?)(9|1[0-9]|2[0-1])e_?(s\.|siècle))/`

[16] Afin de permettre leur rétroconversion éventuelle, cette limite est exceptionnellement étendue à 2999 pour les articles de la refonte de la tranche alphabétique B- en cours de rédaction. L'expression régulière est alors `/(ca\\.?|env\\.?)?([1-2][0-9]{3})/`.

Ordre de détection

Ces quatre formats sont détectés dans l'ordre a-d-c-b, afin de résoudre les cas d'enchâssements d'une expression régulière dans une autre (par exemple, si le siècle qui termine une fourchette était balisé préalablement à celle-ci, la fourchette ne serait pas reconnue).

5.4.15.3 Gestion des ambiguïtés

> Certaines séquences de chiffres ressemblent à des datations mais n'en sont pas. C'est le cas surtout des dates d'édition ou des numéros de page qui peuvent suivre une référence bibliographique.

Ces ambiguïtés sont résolues par le balisage préalable des références bibliographiques (cf. *tag-biblio*) et la création d'une chaîne virtuelle qui les rend invisibles lors de la détection des dates. Par ailleurs, les dates de moins de quatre chiffres ne sont pas balisées comme dates, ce qui génère certes quelques rares faux négatifs, mais évite surtout le balisage erroné de nombreux faux positifs.

5.4.15.4 Résumé

Dépendances

L'algorithme *tag-date* dépend de l'algorithme *tag-biblio* et *tag-renvoi*.

Description

Tag-date n'est pas appliqué aux articles de renvoi. Dans les autres articles, il identifie et balise, dans chaque paragraphe (y compris les notes) sauf dans l'entrée, les mentions de datations. Celles-ci sont reconnues grâce à plusieurs expressions régulières qui dépendent de la métalangue de l'article. L'ordre de recherche des expressions régulières est défini de façon à éviter les ambiguïtés qui pourraient survenir en cas de texte correspondant à plus d'une expression régulière.

La recherche d'une expression régulière s'effectue dans une chaîne virtuelle, dans laquelle la majorité des balises sont transparentes et certaines invisibles (cf. partition). La chaîne virtuelle est recréée autant de fois qu'il y a d'expressions régulières à rechercher, afin que les <dates> précédemment balisées soient rendues invisibles.

Le texte reconnu par les expressions régulières est balisé <date>.

5.4.16 Balisage des catégories grammaticales : tag-gram

1. a. α. Apr. <i>preterir</i> v. n. „passer" (ca. 1300)<e>1</e>) ; mfr. „mourir" (ca. 1508) ; v. a. „laisser tomber dans l'oubli" (1530).	1. a. α. Apr. <i>preterir</i> <gram>v. n.</gram> „passer" (ca. 1300)<e>1</e>) ; mfr. „mourir" (ca. 1508) ; <gram>v. a.</gram> „laisser tomber dans l'oubli" (1530).

(FEW 9, 322b, PRAETERIRE 1 a α)

5.4.16.1 Objectifs de l'algorithme

L'algorithme *tag-gram* a pour objectif de reconnaître la molécule de l'infrastructure qui concerne les catégories grammaticales et de la baliser comme suit :

<gram>m.</gram>

Le balisage des catégories grammaticales se justifie indirectement, dans la mesure où il participe à la reconnaissance des cellules lexicales (cf. *tag-unit*).

5.4.16.2 Critères de détection

> Les catégories grammaticales apparaissent après le signifiant, sous forme abrégée. Elles sont toutefois souvent ellipsées. Une liste de ces catégories est présentée dans Matthey et Thibault en préparation.

Une liste *few-gram-base* a été créée, contenant les abréviations grammaticales en allemand et en français.

5.4.16.3 Résolutions des ambiguïtés

La liste *few-gram-base* ne comporte aucun cas de collision avec d'autres listes de mots-clés. Des ambiguïtés sont néanmoins possibles :

> Certaines abréviations grammaticales sont ambiguës, car ces mêmes abréviations, utilisées dans d'autres contextes, ne font pas partie de la molécule de la catégorie grammaticale. C'est le cas de "v.", qui peut être l'abréviation de *verbe*, mais aussi de *voir* (« dp. 1728, Le Clerc, v. Trév 1743 », FEW 25, 882a, AUGMENTATOR) ou de "s.", qui peut signifier *siècle* (dans une date en français) ou *siehe* (avant une référence bibliographique ou dans un renvoi en allemand).

Certaines ambiguïtés sont résolues par le séquençage des algorithmes et par la création dans *tag-gram* d'une chaîne virtuelle qui rend invisibles les éléments susceptibles de contenir des abréviations similaires, tels que les renvois, les dates ou les marqueurs alphanumériques.

Parmi les ambiguïtés qui ne peuvent être résolues par l'invisibilité des éléments d'une chaîne virtuelle se trouve l'abréviation "v.". Cette abréviation est, dans certains cas, désambiguïsée par l'algorithme *tag-precisions*, qui vient après *tag-gram* dans la séquence.

Les autres ambiguïtés ne sont pas résolues par les algorithmes. De faux positifs (abréviations balisées <gram> alors qu'elles n'expriment pas une catégorie grammaticale) peuvent donc apparaître dans le document résultant de l'application de *tag-gram*. Ces erreurs de balisage n'ont aucune conséquence négative sur la suite du processus de rétroconversion.

5.4.16.4 Détection de catégories grammaticales situées hors molécule

L'indication de la catégorie grammaticale se trouve parfois à l'intérieur des guillemets de définition :

Dans le cas des verbes à plusieurs diathèses, l'indication de la catégorie grammaticale peut se trouver à l'intérieur des guillemets, avec la définition : « Vienne *dénâler* „v.a. disjoindre, disloquer, mettre hors d'usage ; v.r. s'user, se disjoindre, se disloquer" » (FEW 25, 580a, *ASSŬLARE [2] I 1). (Matthey et Thibault en préparation)

Dans ce cas, il s'agit bien de la molécule de la catégorie grammaticale, qui a été placée à l'intérieur de la définition pour des raisons économiques de rédaction. Nous avons choisi de ne pas baliser ces catégories grammaticales, car leur position particulière induit des problèmes divers (modification du schéma XML, prise en compte de ce cas particulier dans les algorithmes suivants) alors que leur reconnaissance n'est pas nécessaire à l'identification des cellules lexicales (rappelons que le balisage des catégories grammaticales a pour but premier de permettre cette identification, → 3.6.3.3). Par conséquent, les définitions sont rendues invisibles dans la chaîne virtuelle traitée par *tag-gram*.

Le cas des mots grammaticaux est tout à fait particulier :

> Les articles où sont traités les étymons ayant donné lieu à des mots grammaticaux donnent des renseignements sur la catégorie grammaticale dans leur structure même (v. par ex. s.v. ĬLLE : « I. Demonstrativpronomen [...]. II. Personalpronomen [...]. III. Determinativpronomen [...]. IV. Artikel [...]. V. Demonstrativpronomen (verstärkt) [...]. » ou à l'intérieur des guillemets qui encadrent normalement la définition (« Afr. *il* „pron. accentué 3e pers. m. pl. nom." ») (FEW 4, 551a). (Matthey et Thibault en préparation)

Dans ces cas particuliers, les catégories grammaticales font partie intégrante d'autres types d'information (titre de paragraphe, définition). Elles ne sont dès lors pas à considérer, dans la structure du FEW, comme des molécules de catégorie grammaticale et

ne sont donc pas balisées par *tag-gram*.

> Les formes verbales citées ne sont pas toujours à l'infinitif ; on précise alors le mode, le temps et la personne : « Allier *d ā l* v.n. IndPr 3 „(le foin) sèche" » (FEW 25, 580a, *ASSŬLARE[2] I 1).

Dans certains cas, le rédacteur peut vouloir fournir des renseignements complémentaires sur la conjugaison. Il peut le faire en note (cf. n 1 de l'article ARDĒRE, signé par Gilles Roques, FEW 25, 147a). (Matthey et Thibault en préparation)

Les abréviations grammaticales sont reconnues uniquement lorsqu'elles se présentent sous la forme abrégée conventionnelle qui est utilisée dans les molécules de la cellule lexicale. Les précisions et les renseignements complémentaires ne sont pas reconnus par les algorithmes.

5.4.16.5 Résumé

Dépendances

L'algorithme *tag-gram* dépend de *tag-geoling*, *tag-numbering*, *tag-renvoi*, *tag-def* et *tag-date*.

Description

L'algorithme *tag-gram* identifie et balise, dans l'entrée et dans chaque paragraphe de l'article (y compris les notes), les catégories grammaticales.

Le traitement d'un paragraphe consiste à y rechercher les mots-clés de la liste *few-gram-base*. La recherche est effectuée dans une chaîne virtuelle, dans laquelle la plupart des balises sont transparentes et certaines invisibles (cf. partitions).

5.4.17 Balisage des signifiants : tag-form

— <geoling>Nfr.</geoling> verbe <i>inchoatif</i> <def>„qui indique le commencement de l'action"</def>	— <geoling>Nfr.</geoling> <form>verbe <i>inchoatif</i></form> <def>„qui indique le commencement de l'action"</def>

(FEW 4, 623a, ĬNCHOARE II 2 b)

5.4.17.1 Objectifs de l'algorithme

L'algorithme *tag-form* a pour objectif principal de reconnaître la molécule de l'infrastructure qui concerne les signifiants des lexèmes. Ces signifiants, en italiques dans le FEW, sont balisés comme suit :

<form><i>texte</i></form>

Outre son intérêt direct évident (→ 3.6.3), le balisage des signifiants dans les parties

de documentation présente un intérêt indirect dans la mesure où il participe à la reconnaissance des cellules lexicales (cf. *tag-unit*). Afin de faciliter cette dernière (effectuée par *tag-unit*), *tag-form* doit étendre (dans la documentation uniquement) le balisage des signifiants aux éléments de collocation introduits en romaines :

> <form><i>gérer</i> une fonction, etc.</form> „exercer“ (FEW 4, 119b, GĔRĔRE II 1)

Dans le commentaire et les notes, *tag-form* balise également tout morceau de texte en italiques susceptible d'être une forme. Lorsque l'extrait balisé contient plus d'un mot, *tag-form* ajoute à la balise <form> insérée un attribut *type*, signalant la possibilité qu'il s'agisse, non d'une forme, mais d'une citation :

> <form type="citation ?"><i>texte</i></form>

Tag-form tente également de reconnaître les locutions, signalées par la valeur "locution ?" de l'attribut *type*. Les extraits en alphabet grec sont également reconnus par *tag-form*. Un attribut *lang* indique qu'il s'agit d'un texte en grec :

> <form lang="grec"><i>texte en grec</i></form>

5.4.17.2 Critères de détection des signifiants

> IPo. Les formes (par essence galloromanes) n'apparaissent pas dans le champ de l'entrée, mais bien dans le reste de l'article : documentation, commentaires et notes.
> ITy. Elles sont en italiques.
> ITe. Une petite partie seulement des formes sont répertoriées dans les index du FEW.

Les indicateurs textuels ne sont pas fiables : l'utilisation d'une liste de mots-clés provenant des index du FEW ne permettrait de reconnaître qu'une très petite partie des formes. L'italique, propriété commune à toutes les formes, est en revanche un indicateur assez fiable pour repérer toutes les formes et éviter les faux négatifs. Il est donc tout à fait approprié de se servir de ce critère, à condition que le balisage typographique inséré dans le document lors de la saisie soit correct. Les erreurs du FEW concernant la mise en italiques sont très rares et susceptibles de correction lors de la saisie du texte. Les erreurs de saisie sont quant à elles corrigées par plusieurs algorithmes de prétraitement : *streamline-void-tags*, *detect-dubious-spacing* et *streamline-quotes* corrigent des balises <i> et </i> mal placées.

> Tous les mots en italiques ne sont pas des signifiants. En italiques apparaissent notamment des étymons, des affixes, des phonèmes, des citations et des caractères grecs.

Le critère de la mise en italiques n'est en revanche pas assez spécifique pour éviter les faux positifs. Une partie des ambiguïtés est résolue par le séquençage des algorithmes et la création de chaînes virtuelles. Les étymons par exemple, qui peuvent apparaître en italiques dans le texte des articles, sont balisés avant les formes (cf. *tag-etymon*) et rendus invisibles dans *tag-form*. Il en va de même pour les affixes (cf. *tag-affix*), les marqueurs alphanumériques grecs (cf. *tag-numbering*) et les mots en italiques

apparaissant dans des définitions (cf. *tag-def*).

Afin de ne pas baliser comme formes, dans le commentaire notamment, des graphèmes isolés symbolisant des phonèmes, nous considérons en outre qu'une forme est constituée de plus d'un caractère. Tout élément <i> contenant un seul caractère, contenant deux caractères dont un tiret ou contenant trois caractères dont deux tirets est considéré comme non valable et n'est pas balisé.

Par ailleurs, tout élément <i> contenant deux mots ou plus[17] est balisé, mais considéré provisoirement comme une locution (ajout d'un attribut *type* = "locution ?" à la balise <form>) ou une citation (ajout d'un attribut *type* = "citation ?" à la balise <form>) selon qu'il apparaît respectivement dans la documentation ou dans le commentaire et les notes. Le point d'interrogation qui suit les termes *locution* et *citation* est là pour rappeler qu'il s'agit d'une supposition et non d'une certitude, notamment en raison du partage non univoque entre les deux lieux d'apparition (documentation pour les locutions, commentaire et notes pour les citations)

Une fois ces filtres mis en œuvre, les extraits en italiques retenus peuvent être de trois types : des signifiants, des mots grecs ou des étymons non reconnus par *tag-etymon*.

Les extraits en grec sont facilement détectables par l'indicateur typographique qu'est le codage des caractères. Ils sont balisés <form> avec l'ajout d'un attribut *lang*="grec".

Les étymons non détectés par *tag-etymon* ne sont pas reconnus comme étymons par *tag-form*. La présence d'une balise <lang> dans le contexte immédiat de la forme pourrait permettre une désambiguisation, mais il existe un risque de confusion avec des mentions de formes non galloromanes, elles aussi précédées de balises <lang>. Seule une lecture par un opérateur humain permettra de les désambiguiser.

5.4.17.3 Extension des formes dans la documentation

> ITy. Dans les paragraphes de documentation, les formes constituent une molécule de la cellule lexicale. Or, cette molécule ne se présente pas toujours exclusivement en italiques. Le signifiant qui constitue le centre de la molécule peut être précédé ou suivi d'un mot qui forme avec lui un syntagme, mais qui n'est pas en italiques, parce que le rédacteur, considérant qu'il s'agit d'un collocatif et non pas d'un élément d'une locution, ne l'inclut pas dans la définition qui suit :
>
> <form><i>gérer</i> une fonction, etc.</form> „exercer" (FEW 4, 119b, GĔRĔREII 1)

Par souci de modélisation efficace de la cellule plus que d'exactitude linguistique, nous considérons que l'ensemble du syntagme appartient au composant <form>, même si seul le signifiant du lexème apparaît effectivement en italiques. Cette décision justifie la conservation du balisage typographique <i> en plus du balisage sémantique <form>. Nous appellerons *extension* (du signifiant) le(s) mot(s) en romaines qui accompagne(nt) le signifiant *stricto sensu*.

[17]Par « mot », nous entendons ici plus particulièrement une séquence de caractères délimitée à sa gauche et à sa droite par une espace (outre les balises <i> et </i>.

Une fois reconnu un signifiant, l'algorithme doit dès lors vérifier s'il est précédé ou suivi d'une extension. Cette extension a pour particularités d'être en romaines non grasses et non italiques et de précéder ou suivre directement, c'est-à-dire sans ponctuation, le signifiant en italiques. À partir du signifiant détecté, l'algorithme peut étendre la forme à droite ou à gauche s'il ne rencontre pas de ponctuation et que les caractères trouvés n'appartiennent pas à un autre composant de la cellule, tel qu'une étiquette géolinguistique, une catégorie grammaticale ou un sigle bibliographique (tous éléments qui partagent les mêmes particularités que l'extension et peuvent donc être confondus avec celle-ci). Cette dernière règle nécessite le balisage préalable des étiquettes géolinguistiques (*tag-geoling*), des catégories grammaticales (*tag-gram*) et des sigles bibliographiques (*tag-biblio*).

Les expressions régulières /([A-Za-z]+\\s*)+/ et /(\\s*[A-Za-z]+)+/, recherchées dans une chaîne virtuelle rendant visibles la plupart des balises (dont <geoling>, <gram> et <biblio>), permettent de détecter une extension respectivement à gauche et à droite selon les règles définies. Afin que les caractères spéciaux du FEW n'entravent pas la détection d'une extension, cette expression régulière est recherchée après aplatissement des caractères ($\rightarrow$ 4.2.2).

5.4.17.4 Résumé

Dépendances

L'algorithme *tag-form* dépend de *detect-dubious-spacing*, *streamline-layout-tags*, *tag-etymon*, (*tag-numbering*), (*tag-def*), *tag-geoling*, *tag-gram*, *tag-biblio*, *split-doc-com*, *split-mixt-art*.

Description

Tag-form identifie et balise, dans chaque paragraphe de l'article (y compris les notes) sauf dans l'entrée, les signifiants des lexèmes, définis comme des chaînes de caractères situées entre des balises <i>...</i>, et toute autre section en italiques susceptible d'être une locution ou une citation.

Le traitement de chaque paragraphe s'effectue dans une chaîne virtuelle qui rend transparentes ou invisibles la plupart des balises (cf. partition). Les étymons, les affixes, les définitions et les marqueurs alphanumériques sont notamment rendus invisibles.

Les sections en italiques contenant une seule lettre non grecque (signes de ponctuation non compris) est balisée avec un attribut *type*= " ?". Toute autre section en italique est balisée <form>. Si l'élément <form> est constitué uniquement de lettres grecques (hors ponctuation et espaces), un attribut *lang*="grec" est ajouté à la balise <form>. Sinon, selon le contenu de l'élément <form> et selon qu'on se trouve dans un paragraphe de documentation, de commentaire ou de notes, un attribut *type*, ayant pour valeur "citation ?" ou "locution ?", est ajouté à la balise <form>.

Dans les paragraphes de documentation uniquement (entourés d'une balise <doc>), la section balisée peut être étendue d'un ou de plusieurs mots vers la gauche ou vers la droite. Cette opération d'extension est effectuée uniquement si les contextes gauche et/ou droit répondent à des expressions régulières bien définies (voir ci-dessus). Les

expressions régulières sont recherchées en activant la méthode d'aplatissement des caractères (→ 4.2.2).

5.4.18 Balisage des concepts : tag-concept

<entry>tabac.</entry> Nfr. <i>toquette</i> f. „feuilles de tabac roulées en andouille" Enc. 16, 420b. → toque.	<entry><concept>tabac</concept>. </entry> Nfr. <i>toquette</i>f. „feuilles de tabac roulées en andouille" Enc. 16, 420b. <renvoi>→ <concept>toque</concept> </renvoi>.

(FEW 21, 149a, tabac)

5.4.18.1 Objectifs de l'algorithme

L'algorithme *tag-concept* a pour objectif de baliser les concepts qui apparaissent au dernier niveau du classement onomasiologique dans les volumes 21 à 23, ainsi que ceux qui apparaissent, dans tous les volumes du FEW, en tant que cible d'un renvoi.

Après application de *tag-concept*, un article des volumes 21 à 23 comportera donc au minimum un élément XML <concept>, situé dans le champ de l'entrée. Ce balisage correspond aux décisions prises lors de la modélisation du FEW (→ 3.5.2).

Le balisage des concepts en position de vedette se présente comme suit :

<concept type="vedette">concept</concept>

Les concepts présents comme renvois dans les articles sont balisés comme suit :

<renvoi>→ <concept>concept</concept></renvoi>

5.4.18.2 Détection des concepts-vedettes

Indicateurs de reconnaissance

> ITe. Le classement onomasiologique des volumes 21 à 23 du FEW est connu. Un index des concepts, établi par Yan Greub, a été publié dans le fascicule 160 du FEW.

À partir de ces données, une liste de mots-clés correspondant aux concepts situés au plus bas niveau du classement a pu être constituée, sous le titre de *few-concept-base*.

> IPo/ITy. Les concepts que nous considérons comme vedettes des articles des Inconnus se présentent en caractères gras et constituent le champ de l'entrée de ces articles.

Combinaison des indicateurs

À partir des indicateurs textuels, positionnels et typographiques, il est possible d'écrire une règle de reconnaissance. Un extrait de texte fewien est un concept-vedette s'il respecte trois conditions : il appartient à la liste *few-concept-base*, il se trouve dans le champ de l'entrée (<entry>...</entry>) d'un article des volumes 21 à 23 et il est entouré des balises ....

5.4.18.3 Détection des concepts de renvoi

Indicateurs

La reconnaissance des concepts de renvoi se base sur les mêmes indicateurs textuels que la reconnaissance des concepts-vedette. En revanche, les indicateurs positionnels et typographiques sont différents :

> IPo. Un concept de renvoi peut apparaître n'importe où dans un article, excepté dans le champ de l'entrée. Il est toujours précédé d'une flèche (mais pas forcément directement) :
>
> Bearn. boulumague f. → bugrane. (FEW 21, 148b, lotier)
>
> [...] → noch salsifis des prés (FEW 21, 149a, houblon)
>
> ITy. Un concept de renvoi apparaît en caractères romaines.

Combinaison des indicateurs

La reconnaissance d'un concept de renvoi est possible en combinant les indicateurs textuels et positionnels. Un extrait de texte fewien est un concept de renvoi s'il appartient à la liste *few-concept-base*, s'il se trouve en dehors du champ de l'entrée et s'il est précédé d'une flèche (caractère Unicode U+2192). La flèche est incluse dans le balisage du renvoi.

5.4.18.4 Résolution des collisions

La liste *few-concept-base* comporte des cas de collisions avec les listes *few-bib-base* (1 collision), *few-com-mot-base* (6 collisions), *few-com-wort-base* (1 collision), *few-etymon-base* (64 collisions) et *few-etymon-cache-base* (5 collisions).

Lors du traitement de l'entrée, les collisions avec *few-etymon-base* sont résolues par le fait que la détection des mot-clés de *few-concept-base* a lieu uniquement dans les articles des volumes 21 à 23 (dans lesquels aucun étymon-vedette n'a, normalement, été balisé par *tag-etymon*). Les autres collisions sont résolues par le fait qu'aucun mot-clé des autres listes n'apparaît dans le champ de l'entrée de ces articles.

Lors du traitement des paragraphes, les collisions avec *few-etymon-base* sont résolues par le séquençage des algorithmes (application de *tag-etymon* avant *tag-concept*) et par la création de chaînes virtuelles rendant invisibles, lors de la détection des mots-clés de *few-concept-base*, les éléments <etymon>. La collision avec *few-bib-base* est résolue idéalement en rendant transparentes les balises <biblio>, puis en examinant si l'élément <renvoi> ou <concept> inséré par *tag-concept* contient (ou est directement

inclus dans) des balises <biblio>. Si c'est le cas, les balises <biblio> sont supprimées : en effet, la présence d'une flèche à gauche du mot-clé est un indicateur assez spécifique pour désambiguïser ces cas de façon systématique. Comme la seule collision connue concerne le mot-clé « Vierge », l'algorithme peut être optimisé en effectuant la vérification de présence des balises <biblio> uniquement pour ce mot-clé. Les collisions avec les listes *few-com-mot-base* et *few-com-wort-base* ne posent aucun problème, puisque les mots-clés de ces listes ne font jamais l'objet d'un balisage.

5.4.18.5 Résumé

Dépendances

Tag-concept dépend uniquement de l'algorithme *tag-entry*.

Tag-concept utilise une seule liste de mots-clés : *few-concept-base*.

Description

L'algorithme *tag-concept* identifie et balise, dans les articles des volumes 21 à 23 uniquement, les concepts-vedettes situés dans le champ de l'entrée. Il identifie et balise également, dans tous les articles du FEW, les concepts de renvois. Ces derniers sont détectés dans tous les paragraphes de l'article hors entrée (paragraphes de notes y compris).

Le traitement de l'entrée consiste à y rechercher les mots-clés de *few-concept-base* (après construction d'une chaîne virtuelle rendant transparentes la plupart des balises et rendant inattendue la balise <etymon>) et à vérifier ensuite que le mot-clé détecté se trouve inclus dans des balises .... Le cas échéant, le mot-clé est balisé <concept>...</concept>. Si aucun mot-clé n'est trouvé dans le champ de l'entrée, ou si le mot-clé trouvé n'est pas inclus dans des balises ..., un avertissement est émis (il s'agit en effet d'une situation anormale qui exige vérification).

Le traitement d'un paragraphe consiste à y rechercher les mots-clés de *few-concept-base*, après construction d'une chaîne virtuelle rendant transparentes la plupart des balises (dont <biblio>) et certaines invisibles (dont <etymon> et <appelsnotes>). Les mots-clés détectés sont balisés <concept>...</concept> uniquement s'ils sont précédés d'une flèche (et d'une espace éventuelle). L'ensemble constitué par la flèche et le concept est balisé <renvoi>...</renvoi>. Si aucune flèche n'est trouvée, le mot-clé n'est pas balisé.

5.4.19 Balisage des précisions : tag-precisions

1. a. <geoling>Afr.</geoling> <geoling>mfr.</geoling> <form><i>personel</i></form> <gram>adj.</gram> <def>„qui se conjugue à toutes les personnes (d'un verbe)“</def> (<biblio>GuernesS-Thomas</biblio> ; <date>ca. 1410</date>, <biblio>ZFSL 1, 31</biblio>).	1. a. <geoling>Afr.</geoling> <geoling>mfr.</geoling> <form><i>personel</i></form> <gram>adj.</gram> <def>„qui se conjugue à toutes les personnes (d'un verbe)“</def> <precisions>(<biblio>GuernesS-Thomas</biblio> ; <date>ca. 1410</date>, <biblio>ZFSL 1, 31</biblio>)</precisions>.

(FEW 8, 273a, PERSONALIS)

5.4.19.1 Objectifs de l'algorithme

L'algorithme *tag-precisions* a pour objectif de baliser le composant précisions, qui englobe les molécules dites « facultatives » de référence bibliographique, de localisation, de datation et d'informations complémentaires, situées à la fin de la cellule lexicale (→ 3.6.4).

Outre les raisons linguistiques évoquées plus haut (→ 3.6.1), le regroupement des quatre molécules facultatives de la cellule lexicale en un seul élément précisions présente un intérêt algorithmique pour la reconnaissance des cellules lexicales (cf. *tag-unit*).

Les précisions sont balisées <precisions>, de la façon suivante :

<precisions><biblio>SSimon</biblio></precisions>

5.4.19.2 Détection des précisions situées entre parenthèses

Indicateurs positionnels

> IPo. Les précisions se trouvent souvent entre parenthèses. Toutefois, les parenthèses peuvent (mais très rarement) servir dans le FEW à d'autres fonctions : elles servent notamment à encadrer des informations morphologiques ou syntagmatiques portant sur les signifiants, et elles peuvent apparaître dans les définitions (cf. Matthey et Thibault en préparation) :
> Afr. *oiant* (+ dénomination d'un ensemble de personnes) „(dire qch, parler) en présence de“ (WaceConcA—JCondOisR [...]) (FEW 25, 848a, AUDĪRE 1 2 a)
> En outre, les parenthèses fermantes d'appels de notes peuvent apparaître n'importe où (cf. *tag-appelnote*).

Les parenthèses d'appels de note doivent être désambiguïsées par le balisage préalable des appels de note (cf. *tag-appelnote*). Le balisage préalable des définitions (cf.

tag-def) et des formes (cf. *tag-form*) permet quant à lui de filtrer la plupart des parenthèses qui ne sont pas des parenthèses de précision. Ces désambiguïsations permettent que la présence de parenthèses constitue un indicateur assez spécifique pour pouvoir être utilisé. Il doit cependant être combiné à d'autres critères.

Indicateurs structurels

> IS. Les parenthèses de précisions contiennent des sigles bibliographiques, des localisations, des datations ou des informations complémentaires.

Le balisage préalable des sigles bibliographiques (*tag-biblio*), des localisations (*tag-geoling*) et des datations (*tag-date*) permet d'utiliser ces types d'information comme critères de détection. Il faut toutefois émettre deux réserves. D'une part, les éléments balisés <geoling> sont ambigus et donc inutilisables comme indicateurs, car ils peuvent remplir les deux fonctions d'étiquette géolinguistique ou de localisation, la première étant en outre la plus fréquente. D'autre part, le balisage des références bibliographiques n'est pas totalement fiable, car il dépend d'une liste de mots-clés qui n'est pas nécessairement exhaustive.

> Les precisions peuvent également contenir des unités :
> ang. *maiguerlin* „id.„malingre" (daher Montjean *maiguer-mine* „individu de mine chétive"[4]), maug. *maigre-mine* „individu outrageusement maigre"), MaineL. *m e g e r l ẽ* „maigrelet" ALFSuppl p 435 (FEW 6/1, 6b, MACER 1)

L'apparition ponctuelle de cellules enchâssées, se présentant entre parenthèses comme les précisions, empêche d'utiliser la présence d'éléments <form>, <gram> ou <def> comme indicateurs négatifs (c'est-à-dire des indicateurs dont l'absence serait signifiante).

Combinaison des critères

En combinant les critères positionnels et structurels, il est possible de détecter les précisions avec une grande fiabilité. Dans un paragraphe (ou sous-paragraphe) de documentation, la présence d'éléments <biblio> ou <date> à l'intérieur de parenthèses est un indicateur suffisamment fiable pour reconnaître des précisions. Le balisage des précisions inclut tout ce qui se trouve à l'intérieur des parenthèses.

Afin de remédier au problème de non-exhaustivité du balisage des sigles bibliographiques, des parenthèses ne contenant aucun élément <biblio> ou <date> sont également balisées comme précisions, mais avec un statut "ambigu" pour vérification manuelle.

Lorsque l'élément précisions contient des balises <form>, <gram> ou <def>, un attribut ayant pour valeur "contains-unit" est ajouté à la balise <precisions> afin de permettre une vérification éventuelle.

5.4.19.3 Détection de précisions non entourées de parenthèses

> Quand le composant des précisions se résume à une seule référence bibliographique (éventuellement accompagnée d'une indication de volume, de page, de vers etc.), les parenthèses sont omises :
>
> Bnorm. *houllebiche* „chien de mer“ Seguin 105 (FEW 18, 72a, HOLIBUT 2)

Les sigles bibliographiques et les datations qui ne sont pas entourés de parenthèses seront également considérés comme des précisions à condition qu'ils soient directement précédés d'une molécule <geoling>, <def>, <gram> ou <form>. En effet, toute unité contient obligatoirement un de ces quatre éléments avant la mention éventuelle de précisions.

5.4.19.4 Résolution d'ambiguïtés

Le balisage des précisions permet de lever des ambiguïtés qui n'étaient pas résolvables auparavant.

Vérification des paires de parenthèses

Lors de la détection des précisions situées entre parenthèses, l'algorithme vérifie que les parenthèses (hors parenthèses d'appels de note) des paragraphes traités vont bien par paires, ce qui permet de repérer d'éventuelles erreurs ou incohérences dues au FEW lui-même ou à la saisie. En cas de non parité des parenthèses, un avertissement est émis pour vérification et correction manuelle.

Élimination de faux positifs <gram>

La chaîne de caractères "v." est ambiguë dans le FEW : elle peut signifier "verbe", mais aussi "voir", notamment devant un sigle bibliographique. Or, cette chaîne est balisée <gram> par *tag-gram*. Lorsque l'élément <precisions> balisé englobe un élément <gram> contenant la chaîne "v." et précédant un sigle bibliographique, les balises <gram> sont supprimées.

5.4.19.5 Résumé

Dépendances

La reconnaissance des précisions nécessite la reconnaissance préalable des parties documentaires de l'article (cf. *split-doc-com* et *split-mixt-art*). Il nécessite également le balisage préalable des molécules <geoling>, <form>, <gram>, <def>, <biblio> et <date> (*tag-geoling*, *tag-form*, *tag-gram*, *tag-def*, *tag-biblio*, *tag-date*), ainsi que des appels de note (*tag-appelnote*).

Description

L'algorithme *tag-precisions* détecte et balise (au moyen de l'élément XML <precisions>), dans les paragraphes marqués <doc> ou <mixt> d'un article, le composant de la cellule lexicale appelé "précisions" (→ 3.6.4). Le traitement d'un paragraphe s'effectue en trois phases successives : tout d'abord le balisage des sections entre parenthèses, ensuite le balisage d'éléments <biblio> ou <date> non entourés de parenthèses, enfin la désambiguïsation des éléments <gram> qui se trouveraient dans les éléments <precisions> balisés.

Le balisage de chaque section située entre des parenthèses de haut niveau (c'est-à-dire non incluses dans d'autres parenthèses) s'effectue après vérification de la bonne imbrication des parenthèses dans le paragraphe. Ce premier traitement a lieu dans une chaîne virtuelle qui rend transparentes la plupart des balises et invisibles d'autres comme les appels de note. Chaque élément <precisions> balisé reçoit un attribut dont la valeur est "contains-unit" s'il contient au moins un élément <geoling> et un élément <form>[18], "ok-wraps-parentheses" s'il contient au moins un élément <biblio> ou <date>, et "ambiguous" sinon.

Le balisage d'éléments <biblio> ou <date> non entourés de parenthèses s'effectue en itérant sur tous les <biblio> et <date> qui ne sont pas dans un élément <precisions> préalablement balisé[19] et en vérifiant que la balise fermante la plus proche à leur gauche est </def>, </form>, </gram> ou </geoling>. Cette vérification s'effectue dans une chaîne virtuelle dans laquelle la plupart des balises sont visibles. Le balisage des éléments <biblio> ou <date> qui répondent à ces critères inclut le texte éventuellement situé à gauche de ces éléments et à droite de la balise fermante </def>, </form>, </gram> ou </geoling>. Chaque élément <precisions> ainsi balisé reçoit un attribut *status* dont la valeur est "ok-wraps-tag".

L'élimination des faux éléments <gram> consiste à supprimer les balises <gram> ... </gram> dont le contenu commence par v., qui seraient situées immédiatement à gauche d'une balise <biblio>.

[18] Une règle plus juste aurait été « s'il contient au moins deux molécules obligatoires (parmi les quatre possibles) qui se suivent », mais en pratique, cette règle est plus complexe à implémenter, alors que la première règle suffit à reconnaître la majorité des cas d'inclusion d'une cellule dans les précisions.

[19] L'implémentation de l'algorithme en Java est plus complexe, car elle vérifie aussi que <biblio> et <date> ne se trouvent pas dans un élément <geoling>, <form>, <gram> ou <def>, suivant un principe de programmation défensive.

5.4.20 Balisage des attestations : tag-attestation

<precisions>(<biblio>Joseph</biblio>– <date>1563</date>, <biblio>Gdf</biblio> ; <biblio>BenSMh</biblio> ; <date>1250</date>, <biblio>DC</biblio>)</precisions>	<precisions>(<attestation><biblio>Joseph</biblio>–<date> 1563</date>, <biblio>Gdf</biblio></attestation> ; <attestation> <biblio>BenSMh</biblio></attestation> ; <attestation> <date>1250</date>, <biblio>DC</biblio></attestation>)</precisions>

(FEW 13/1, 181a, TEMPORALIS I 1)

5.4.20.1 Objectifs de l'algorithme

L'algorithme *tag-attestation* a pour objectif de regrouper, dans les précisions de chaque cellule lexicale d'un article, les molécules qui, ensemble, définissent une période pendant laquelle le lexème est attesté. Conformément à la modélisation proposée, ces molécules sont regroupées au moyen de l'élément XML <attestation> (→ 3.6.4).

5.4.20.2 Indicateurs de reconnaissance

> IPo. Au sein des précisions, les molécules sont hiérarchisées par les signes de ponctuation. Les éléments appartenant à une même fourchette de datation sont séparés par un tiret (court, semi-cadratin ou cadratin). Une source associée à une datation est séparée de cette dernière par une virgule. Le point-virgule sépare des datations différentes :
>
> > Mfr. frm. augmentateur m. “celui qui accroît la valeur, l'importance, l'extension, la puissance (de qch)” (1396, PhMézSust 64 ; 1453, JChartier ChronCharlesVII, tous deux DocDMF ; Mist ; 1504—1611, Hu ; GdfC ; 1505, Molin-Faictz ; Gringore 113 ; Est 1539—Pom 1715 ; ‘rare’ Rob 1988) (FEW 25, 882a, AUGMENTATOR)

Le point-virgule constitue un indicateur à la fois assez fiable et assez spécifique pour délimiter les périodes de datation.

5.4.20.3 Résumé

Dépendances

Tag-attestation dépend uniquement de *tag-precisions* et n'utilise aucune liste de mots-clés.

Description

Tag-attestation identifie et balise, dans les éléments <precisions> d'un article, les attestations (groupes de molécules qui définissent ensemble une période pendant laquelle le lexème est attesté).

Seuls les paragraphes de documentation (marqués <doc> ou <mixt>) sont traités. Dans chacun de ces paragraphes, l'algorithme itère sur les éléments <precisions> et vérifie la valeur de l'attribut *status*. Seuls les éléments <precisions> contenant des parenthèses (valeur "ok-wraps-parentheses") ou ambigus (valeur "ambiguous") sont traités.

Le traitement d'un élément <precisions> s'effectue dans une chaîne virtuelle rendant invisibles la plupart des balises (cf. partition). Dans la chaîne virtuelle ainsi créée, les sections séparées par un point-virgule sont balisées chacune <attestation>.

5.4.21 Balisage des titres : tag-title

<p>Ablt. — <geoling>Mars.</geoling> <form><i>fayoux mavounens</i></form> <def>„haricots sans fils, qu'on mange vert“</def> Avril	<p><title>Ablt.</title> —<geoling>Mars.</geoling> <form><i>fayoux mavounens</i></form> <def>„haricots sans fils, qu'on mange vert“</def> Avril

(FEW 6/1, 52a, MAHÓN)

5.4.21.1 Objectifs de l'algorithme

L'algorithme *tag-title* a pour objectif de reconnaître et de baliser, dans les parties documentaires d'un article du FEW, les marqueurs textuels qui explicitent certains critères de regroupement entre lexèmes ($\rightarrow$ 3.7.2). Il ne balise pas les groupements de lexèmes eux-mêmes, ces derniers étant traités par un autre algorithme (*tag-microstructure*).

Les marqueurs textuels identifiés sont balisés <title>…</title>, conformément à la modélisation proposée ($\rightarrow$ 3.7.2).

5.4.21.2 Indicateurs de reconnaissance

La reconnaissance des marqueurs textuels pose problème autant au niveau positionnel et textuel que typographique.

> IPo. Les marqueurs textuels peuvent apparaître à plusieurs endroits dans le texte du FEW : en début de paragraphe, mais aussi au milieu d'un paragraphe. En milieu de paragraphe, ils sont précédés d'un tiret, d'un point ou d'un point-virgule (ces signes de ponctuation ne sont toutefois pas réservés à ce seul usage) :
>
> lim. *amadra* „(pain) sans levain" DD. **Sekundär** Toulouse *s'amazerá* „se rabougrir, se sécher" (FEW 6/1, 8a, MACERARE I 1 a)[20]
>
> ITe. Les marqueurs textuels constituent un texte non formalisé. Seuls certains marqueurs fréquents tels que « ablt. », « zuss. » ou « redensarten » se présentent souvent sous une même forme (généralement abrégée).
> ITy. Les marqueurs textuels peuvent se présenter sous des formes typographiques très différentes : grasses (« I. 1. a. **Kneten**. », FEW 6/1, 8a, MACERARE), entre crochets (surtout dans la refonte) ou sans caractéristiques typographiques particulières. Les grasses et les crochets ne sont toutefois pas réservés à ce seul usage.

Les critères positionnels, textuels et typographiques sont inefficients individuellement : les signes de ponctuation, les crochets et la mise en grasses ne constituent pas des indicateurs assez spécifiques, puisqu'ils apparaissent dans d'autres positions ; les abréviations connues comme « ablt. » ne sont pas fiables, puisqu'elles ne représentent qu'une partie des marqueurs possibles.

Il est toutefois possible de prendre en compte tous ces indicateurs de façon à repérer un maximum de marqueurs.

5.4.21.3 Règles de reconnaissance

À partir des indicateurs évoqués ci-dessus, nous pouvons écrire quatre règles qui permettent chacune d'identifier de façon fiable une partie des marqueurs textuels d'un article. Ces quatre règles sont appliquées uniquement dans les parties de documentation : elles nécessitent donc la reconnaissance préalable des parties documentaires de l'article (cf. *split-doc-com*).

1. Est considéré comme marqueur textuel un mot appartenant à une liste déterminée de mots-clés. Cette règle permet d'identifier les marqueurs connus, tels que « ablt. », « zuss. », « redensarten », « übertragen » etc., quelle que soit leur position dans le paragraphe. La liste de mots-clés dépend de la métalangue de l'article (français ou allemand) et nécessite donc la reconnaissance préalable de cette métalangue (cf. *tag-signature*). Deux listes de mots-clés ont été constituées : *few-titel-base* (pour les articles en allemand) et *few-titre-base* (pour les articles en français).
2. Est considérée comme marqueur textuel toute séquence de caractères apparaissant juste avant la première étiquette géolinguistique du paragraphe, après le

[20] C'est nous qui mettons en gras.

marquage alphanumérique éventuel. Cette règle permet notamment d'identifier, grâce à leur position en tête de paragraphe, les marqueurs non connus (c'est-à-dire non consignés dans *few-titel-base* et *few-titre-base*). Les particularités typographiques (grasses par exemple) des marqueurs n'empêchent pas leur détection. En revanche, cette règle nécessite le balisage préalable du marquage alphanumérique et des étiquettes géolinguistiques (cf. *tag-numbering* et *tag-geoling*).

3. Est considérée comme marqueur textuel toute séquence de caractères apparaissant au sein d'un paragraphe, entre un signe de ponctuation fort et une étiquette géolinguistique. Seuls le point, le tiret cadratin et le tiret semi-cadratin sont considérés comme des signes de ponctuation forts.

4. Est considérée comme marqueur textuel toute séquence de caractères apparaissant entre crochets qui se trouve en dehors d'une cellule lexicale. Cette règle est surtout intéressante dans les articles de la refonte (volumes 24 et 25), mais peut être appliquée ailleurs également. Elle permet notamment de baliser comme titres les affixes qui, dans la refonte, se trouvent en tête d'un groupement de formes.

La dernière règle nécessite que les affixes préalablement balisés par *tag-affix* soient rendus invisibles lors de la détection des titres. Une deuxième condition d'application de cette règle est le balisage préalable des cellules (cf. *tag-unit*). Un problème se pose ici : la reconnaissance des cellules, effectuée par *tag-unit*, dépend en partie de la reconnaissance des marqueurs textuels effectuée par *tag-title*. On se trouve donc en présence d'un problème de circularité (*tag-title* dépendant de *tag-unit* et vice-versa). Une solution consiste, dans la règle énoncée ci-dessus, à remplacer la condition « hors cellule » par la condition « hors composants d'une cellule ». Cette condition est moins précise, mais opérante en pratique. Le balisage préalable de ces composants, et plus particulièrement des définitions (*tag-def*) et des précisions (*tag-precisions*), est dès lors nécessaire. Comme le balisage de ces composants ne dépend pas de *tag-title*, la circularité est évitée.

5.4.21.4 Combinaison des règles

Parmi les quatre règles énoncées ci-dessus, la première peut empêcher les deuxième et troisième de fonctionner. Dans l'extrait suivant par exemple,

> <p>Auf fische übertragen. <geoling>Nfr.</geoling> <form><i>aigle</i></form>
> <def>„raie des mers d'Europe"</def> (FEW 25, 72b, AQUILA I)

le marqueur *Auf fische übertragen* sera balisé une première fois « Auf fische <titre> übertragen</titre> » par la première règle, puis une deuxième fois « <titre>Auf fische </titre> <titre>übertragen</titre> » par la deuxième règle.

Pour éviter ce problème, il est nécessaire d'appliquer les règles numéro 2 et 3 avant la règle numéro 1. La règle numéro 4 est appliquée au début pour permettre la détection, dans la refonte, de plusieurs titres successifs dont le deuxième serait un affixe, comme dans l'exemple suivant :

> <p><title>Dérivés : </title><title>[+<affix type="suffix"> -IA</affix>]</title>

Chaque détection s'effectue dans une chaîne virtuelle qui rend invisibles les éléments <titre>…</titre> balisés précédemment, afin qu'ils ne soient pas balisés plu-

sieurs fois.

5.4.21.5 Résumé

Dépendances

Tag-title dépend de *split-doc-com*, *tag-numbering*, *tag-geoling*, *tag-form*, *tag-def*, *tag-gram*, *tag-precisions*, *tag-affix* et *tag-signature*.

Tag-title utilise deux listes de mots-clés : *few-titel-base* et *few-titre-base*.

Description

L'algorithme *tag-title* détecte et balise les marqueurs textuels de structuration qui apparaissent dans les parties documentaires d'un article du FEW. Seuls les paragraphes de documentation sont traités.

Le traitement d'un paragraphe s'effectue en quatre phases : (1) détection et balisage des titres entre crochets, (2) détection et balisage des titres de paragraphes, (3) détection et balisage des titres internes, (4) détection et balisage des marqueurs de titres connus. Chacune des phases s'effectue dans une chaîne virtuelle qui rend invisibles les éléments balisés <title> lors des phases précédentes.

La détection des titres entre crochets consiste à chercher les chaînes de caractères entre crochets qui se trouvent en dehors des éléments <geoling>, <form>, <def>, <gram>, <precisions>, <biblio> et <date>. La détection se fait dans une chaîne virtuelle qui rend visibles ces sept balises et qui rend transparentes les balises <affix>, afin que les affixes entre crochets puissent être balisés comme des titres.

La détection des titres de paragraphes consiste à chercher un « caractère délimiteur de titre » en tête de paragraphe, entre la balise <pnum> (si existante) et le premier <geoling> du paragraphe. Un caractère délimiteur de titre est un point, un point-virgule, une virgule, un tiret ou un tiret (semi-)cadratin. La recherche se fait de gauche à droite dans une chaîne virtuelle qui rend visibles la plupart des balises (cf. partition). Dès qu'un caractère délimiteur est détecté, la recherche s'arrête (la suite de la chaîne virtuelle n'est pas traitée). Le texte qui précède le caractère délimiteur est balisé.

La détection des titres internes consiste à chercher un caractère délimiteur de titre dans la section de texte non balisée apparaissant éventuellement juste avant chaque <geoling>. La recherche se fait de droite à gauche dans une chaîne virtuelle rendant visibles la plupart des balises. En cas de détection d'un caractère délimiteur de titre, la section de texte située entre ce caractère et la balise <geoling> est balisée <title>.

La détection des marqueurs connus consiste à chercher les mots-clés de la liste *few-titel-base* (ou *few-titre-base*, si la valeur de l'attribut *lang* de <art> est « french » au lieu de « german »). La recherche des mots-clés se fait en dehors des éléments déjà reconnus et balisés par les algorithmes précédents (dont les éléments <def>, <geoling>, <form>, <gram>, <lang> et <pnum>). La ponctuation forte qui suit éventuellement le marqueur est incluse dans le balisage.

5.4.22 Balisage des cellules lexicales : tag-unit

<geoling>Wallon.</geoling> <form><i>trî</i></form> <def>„friche"</def>, <geoling>Faymonv.</geoling> <geoling>Stavelot</geoling> <def>„terrain non cultivé servant de pâture"</def>	<unit> <geoling>Wallon.</geoling> <form><i>trî</i></form> <def>„friche"</def></unit>, <unit><geoling>Faymonv.</geoling> <geoling>Stavelot</geoling> <imp contents="trî" type="form"/> <def>„terrain non cultivé servant de pâture"</def></unit>

(FEW 3, 158b, DRIESC)

5.4.22.1 Objectifs de l'algorithme

L'algorithme *tag-unit* a pour objectif d'identifier les cellules lexicales dans un article du FEW. Cette identification comporte deux opérations : d'une part, la reconnaissance des limites (début et fin) de chaque mention d'une cellule au sein du texte du FEW ; d'autre part, le rétablissement, au sein de chaque cellule, des molécules ellipsées.

Conformément à la modélisation proposée (→ 3.6.1), la reconnaissance des limites d'une cellule s'effectue par l'insertion des balises <unit> (balise ouvrante, marquant le début de la cellule) et </unit> (balise fermante, marquant la fin de la cellule). Le rétablissement d'une molécule ellipsée s'effectue par l'insertion au sein de la cellule, à l'endroit où devrait se trouver la molécule ellipsée, d'une balise vide <imp/>. Cette balise vide contient deux attributs *type* et *contents*, ayant respectivement pour valeur le nom de la molécule ellipsée (geoling, form, gram ou def) et le contenu ellipsé (→ 3.6.3).

5.4.22.2 Détection des limites de cellule

La première question qui se pose dans la mise au point de l'algorithme *tag-unit* consiste à déterminer comment reconnaître les limites d'une cellule lexicale.

Deux indicateurs sont à notre disposition : d'une part le contenu de la cellule, c'est-à-dire les éléments qui la composent, d'autre part la typographie, c'est-à-dire la ponctuation qui marque une frontière entre les cellules.

Selon la modélisation proposée plus haut, une cellule est constituée de cinq composants, que nous appellerons ici « molécules » : l'étiquette géolinguistique (<geoling>), le signifiant (<form>), la catégorie grammaticale (<gram>), le signifié (<def>) et les précisions (<precisions>). Ces molécules se suivant toujours dans cet ordre, il serait envisageable d'insérer un début de cellule <unit> avant chaque balise <geoling> et une fin de cellule </unit> après chaque composant précisions. Cette règle n'est cependant pas valable, car elle ne tient pas compte, ni du phénomène de l'ellipse qui touche les quatre premières molécules (→ 3.6.3), ni du caractère facultatif des précisions. En structure de surface, aucun des cinq composants d'une cellule n'est obligatoire. En revanche, une cellule doit obligatoirement contenir au moins un des quatre premiers composants, ce

qui constitue, avec l'ordre fixe des composants, un critère fiable permettant l'écriture d'une règle de détection.

Cette règle de détection se fonde sur la succession des molécules pour déterminer le début et la fin d'une cellule. Les molécules reçoivent un numéro d'ordre, de 1 à 5, selon leur type :

1. <geoling>

2. <form>

3. <gram>

4. <def>

5. <precisions>

Les propriétés suivantes sont en outre définies :

1. il ne peut y avoir, au sein d'une même cellule, plus d'une molécule de même niveau ;

2. au sein d'une même cellule, une molécule peut être suivie à sa droite d'une molécule de niveau supérieur uniquement.

Ces deux propriétés sont valables pour toute molécule excepté pour la molécule <geoling>, puisque plusieurs étiquettes géolinguistiques peuvent se succéder au sein d'une même cellule (à condition que seule une espace les sépare) :

> I. 1. Mfr. nfr. *casuel* „fortuit, subordonné à certains cas" (FEW 2, 479b, CASUALIS I 1)
> ang. saint. SeudreS. centr. Schweiz *accomparer* (FEW 2, 970b, COMPARARE I)

Ces propriétés étant définies, la détection d'une cellule peut s'effectuer en itérant de gauche à droite sur les molécules (préalablement balisées) qui apparaissent dans le texte du FEW. La règle de détection est la suivante : à chaque fois qu'une molécule de niveau égal (<geoling> excepté) ou inférieur à la molécule précédente apparaît, elle constitue le début d'une nouvelle cellule.

La typographie constitue un indicateur supplémentaire qui permet de limiter la portée de la règle de détection. Tout d'abord, le paragraphe constitue l'entité maximale d'application de cette règle, c'est-à-dire que cette dernière s'applique au sein d'un même paragraphe uniquement. Ensuite, les signes de ponctuation que sont le point et le tiret marquent une frontière explicite entre les cellules, qui peut être utilisée comme critère de détection également. Le début de paragraphe ou un signe de ponctuation fort (point ou tiret) marquent dès lors obligatoirement le début d'une nouvelle cellule, qui doit impérativement débuter par une étiquette géolinguistique <geoling> (si cette dernière propriété n'est pas respectée, les molécules suivantes ne sont pas traitées, jusqu'à la détection d'une nouvelle molécule <geoling>).

5.4.22.3 Résolution de l'implicite

La seconde question qui se pose dans la mise au point de *tag-unit* est de savoir comment rétablir, au sein de chaque cellule, les molécules implicites.

L'ellipse d'une molécule dans une cellule est un phénomène qui dépend, non de la succession des molécules, mais de la succession des cellules : une molécule est ellipsée lorsque son contenu est déjà cité dans la cellule précédente. Plusieurs cellules peuvent ainsi se succéder en ellipsant la même molécule. Dans l'exemple ci-dessous, la molécule de la définition, absente des cellules saint. St-Seurin *compiet* Minot *conpiet* et npr. *coumplèt*, est à reprendre de la première cellule Mfr. nfr. *complet* :

> Mfr. nfr. *complet* „à quoi il ne manque aucune des parties nécessaires“ (seit ca. 1300, Monstr ; Rhlitt 6, 464), saint. St-Seurin *compiet*, Minot *conpiet*, npr. *coumplèt*. (FEW 2, 982b, COMPLETUS I 1 a)

À partir du moment où les limites des cellules sont connues, il est donc possible de résoudre une molécule implicite en allant rechercher le contenu de cette molécule dans les cellules précédentes. De façon à exploiter l'itération, ce rétablissement de l'implicite est effectué pour chaque cellule au fur et à mesure du parcours (effectué de gauche à droite dans le texte du FEW), en conservant en mémoire (dans un tableau ou toute autre structure de données appropriée) le contenu des quatre premiers niveaux moléculaires et en le mettant à jour à chaque fois qu'une cellule a été traitée. De cette façon, seules les molécules de la cellule précédant la cellule courante doivent être mémorisées.

De nouveau, la typographie joue un rôle important en permettant de délimiter la portée de cette règle. En effet, l'apparition d'un point ou d'un tiret entre deux cellules indique que les informations de la première cellule ne sont pas à reprendre dans la seconde. Dès lors, les données du tableau mémorisant le contenu des molécules doivent être effacées à chaque fois qu'une ponctuation forte est rencontrée entre les cellules.

5.4.22.4 Cas particulier d'implicite : « id »

Dans certains cas (cf. Matthey et Thibault en préparation), une cellule lexicale peut contenir la mention id. (pour idem) en position de molécule. Cette mention reprend généralement à la fois le signifiant et le signifié de la cellule précédente.

Ce cas particulier d'implicite (relatif) ne pose aucun problème, puisque la mention *id.* n'est pas balisée par les agorithmes de reconnaissance *tag-form* et *tag-def*. Elle n'empêche donc pas la résolution de l'implicite suivant les règles édictées ci-dessus : *tag-unit* ignorera simplement cette mention et résoudra l'implicite comme attendu.

5.4.22.5 Résumé

Dépendances

Tag-unit dépend de *split-doc-com*, *tag-geoling*, *tag-form*, *tag-gram*, *tag-def*, *tag-precisions*, *tag-title* et *tag-numbering*.

Description

Tag-unit balise les cellules lexicales d'un article du FEW et rétablit dans chaque cellule les molécules ellipsées. Après identification de toutes les cellules d'un article, *tag-unit* en construit une liste exportable, présentée en fin de document sous forme de commentaire XML.

Une cellule est définie comme une séquence d'une ou de plusieurs molécules. Une molécule est un des cinq éléments XML <geoling>, <form>, <gram>, <def> et <precisions>. À chacun de ces cinq éléments est associé, dans cet ordre, un niveau de 1 à 5. Au sein d'une même cellule, les molécules respectent les deux propriétés suivantes : (1) excepté pour les molécules de niveau 1 (<geoling>), qui peuvent apparaître plusieurs fois, il ne peut y avoir plus d'une molécule de même niveau ; (2) une molécule peut être suivie à sa droite d'une molécule de niveau supérieur uniquement.

Seuls les paragraphes contenant de la documentation (paragraphes <doc> et <mixt>) sont traités par *tag-unit*. Le regroupement des molécules en cellules, le rétablissement des molécules ellipsées (au moyen de l'élément XML <imp/>) et le balisage des cellules (au moyen de l'élément XML <unit>...</unit>) s'effectuent en un seul parcours du paragraphe, effectué de gauche à droite.

Le regroupement des molécules consiste à créer une nouvelle cellule (à partir de la molécule située à droite de la balise fermante </unit> précédemment insérée, ou, si aucun élément <unit> n'a encore été créé, à partir de la molécule située à droite de la balise <p> marquant le début du paragraphe) à chaque fois qu'une des trois situations suivantes est rencontrée :

– un délimiteur de cellule explicite est rencontré : c'est le cas lorsque se trouve entre deux molécules un signe de ponctuation (point ou tiret) ou une balise (<title>) qui marque une frontière entre deux cellules ;

– un délimiteur de cellule implicite est rencontré : c'est le cas lorsqu'une molécule de niveau inférieur ou égal suit une molécule de niveau supérieur (sauf s'il s'agit d'une molécule <geoling>) ;

– la fin de la liste des molécules est atteinte.

Le rétablissement des molécules ellipsées s'effectue au moment où une cellule est créée. Les valeurs des molécules absentes dans la cellule sont déduites à partir des molécules contenues dans la cellule précédente. Le rétablissement de l'implicite se fait uniquement si la cellule précédente ne se termine pas par un point ou un tiret et qu'il existe au moins une molécule explicite <geoling>, <form>, <gram> ou <def> dans la cellule courante. Rétablir l'implicite consiste à itérer sur les « trous » dans la cellule courante et, pour chaque molécule absente, à ajouter une molécule implicite avec le contenu (si existant) de la molécule correspondante dans la cellule précédente. Une fois l'implicite rétabli, le tableau (ou toute autre strucuture de données choisie) gardant en mémoire les valeurs de chaque molécule est mis à jour avec les valeurs de la cellule courante.

Le balisage des cellules consiste à insérer les balises <unit> </unit> aux extrémités de chaque cellule créée et à insérer une balise <imp/> à chaque endroit où une molécule implicite a été détectée. Chaque balise <imp/> reçoit deux attributs : *type*, ayant pour

valeur le nom de la molécule explicitée, et *contents*, ayant pour valeur le contenu de la molécule explicitée.

5.4.23 Réunification des paragraphes : merge-mixt-p

<mixt mixt-type="mixt_doc"> Fr. <i>appellatif</i> adj. „(nom) qui convient à toute une espèce et non à un individu seul“ (ca. 1350–Ac 1878, GlC ; ‚on dit plutôt <i>nom commun’</i> Ac 1932, mfr. <i>appellatif</i> m. „nom commun“ (ca. 1550). </mixt> <mixt mixt-type="mixt_com"> — Lehnwort. </mixt>	<mixt> Fr. <i>appellatif</i> adj. „(nom) qui convient à toute une espèce et non à un individu seul“ (ca. 1350–Ac 1878, GlC ; ‚on dit plutôt <i>nom commun’</i> Ac 1932, mfr. <i>appellatif</i> m. „nom commun“ (ca. 1550). <part_com>—Lehnwort.</part_com> </mixt>

(FEW 25, 32b, APPELLATIVUS)

5.4.23.1 Objectifs de l'algorithme

L'algorithme *merge-mixt-p* a pour objectif de réunifier, dans une unité de traitement (c'est-à-dire un article du FEW), les paragraphes mixtes dont la partie documentaire et la partie consacrée au commentaire ont été séparées en deux sous-paragraphes distincts par l'algorithme *split-mixt-art* (appliqué plus tôt dans la séquence de rétroconversion).

La séparation de ces deux parties par *split-mixt-art* se justifiait, car elle était nécessaire à l'application d'algorithmes tels que *tag-unit* (cf. *split-mixt-art*). Leur réunification est quant à elle essentielle à la reconnaissance de la microstructure de l'article (cf. *tag-microstructure*).

Conformément à la modélisation proposée, la partie commentaire doit être réintégrée au sein du paragraphe et balisée <part-com> (→ 3.7.2).

5.4.23.2 Indicateurs

La présence dans l'article de deux paragraphes <mixt> successifs, dont le premier contient l'attribut *type*="mixt_doc" et le second l'attribut *type*="mixt_com", est un indice suffisant pour reconnaître un paragraphe mixte séparé par *split-mixt-art*.

Lorsque deux paragraphes successifs répondant à ce schéma sont détectés, le second paragraphe doit être considéré comme une partie de commentaire, à réinsérer à la suite du texte du premier paragraphe et à baliser <part_com>…</part_com>.

5.4.23.3 Résumé

Dépendances

L'algorithme *merge-mixt-p* dépend de *split-mixt-art* uniquement.

Description

L'algorithme *merge-mixt-p* itère sur tous les paragraphes d'un article et détecte les successions de paragraphes <mixt> répondant au schéma suivant :

```
<mixt type="mixt_doc"><p>texte du premier paragraphe</p></mixt>
<mixt type="mixt_com"><p>texte du second paragraphe</p></mixt>
```

Le traitement de deux paragraphes successifs répondant à ce schéma consiste à supprimer les balises </p></mixt><mixt type="mixt_com"><p> situées entre les deux paragraphes et à les remplacer par la balise <part_com>. Une balise </part_com> est insérée avant la balise </p> marquant la fin du paragraphe. L'attribut *type*="mixt_doc" du premier paragraphe est également supprimé. Le résultat de l'algorithme est le suivant :

```
<mixt>
<p>texte du premier paragraphe
<part_com>texte du second paragraphe </part_com>
</p></mixt>
```

L'algorithme lui-même ne présentant pas d'intérêt du point de vue linguistique et dépendant fortement de l'implémentation (outils de traitement des balises XML), le code n'est pas exposé dans cette thèse.

5.4.24 Structuration de la documentation : tag-microstructure

```
< !– article map
1 Fr. futur adj. „qui sera“ [. . .]
1 Ablt. Nfr. futurition „caractère de ce qui est futur“ [. . .]
1 Zuss. Mfr. nfr. examen à futur „enquête par avance sur des faits dont on ne veut pas
laisser disparaître les preuves“ [. . .]
2 Nfr. futurisme „mouvement littéraire et artistique créé en Italie par de jeunes écrivains et
artistes, en vue d'une insurrection contre l'académisme“ [. . .]
–>
```

(FEW 3, 929ab, FUTURUS)

5.4.24.1 Objectifs de l'algorithme

L'algorithme *tag-microstructure* a pour objectif d'identifier les groupements de lexèmes dans le champ documentaire d'un article. En effet, le balisage des marqueurs alpha-

numériques (*tag-numbering*) et textuels (*tag-title*) n'est pas suffisant pour situer un lexème dans la structure de l'article : il faut également mettre les lexèmes en relation avec les marqueurs, ce qui revient à identifier la portée de ces derniers.

Conformément aux choix effectués lors de la modélisation (→ 3.7.2), cet objectif revient, d'une part, à situer chaque paragraphe dans la structure globale de l'article et, d'autre part, à identifier les groupements de lexèmes les plus importants au sein de chaque paragraphe. *Tag-microstructure* effectue donc deux balisages :

1. Le premier consiste à insérer, au sein de chaque balise <p>, un attribut *structid* qui situe le paragraphe au sein de la structure alphanumérique de l'article ;
2. Le second consiste à insérer, au sein de chaque paragraphe, des balises <group> ...</group> autour de chaque groupement de lexèmes explicité par un marqueur textuel.

Par ailleurs, *tag-microstructure* présente, sous forme de commentaire XML placé en début d'article, un plan reprenant le marquage alphanumérique de l'article et la première cellule de chaque paragraphe, de façon à donner au lecteur une vision synthétique et approximative de la structure de l'article.

5.4.24.2 Première phase : identification structurelle des paragraphes

L'identification structurelle de chaque paragraphe est possible uniquement si l'article traité structure la documentation au moyen d'un marquage alphanumérique. Le cas échéant, la reconnaissance de ce marquage alphanumérique a été effectuée préalablement par l'algorithme *tag-numbering* au moyen d'une balise <pnum> (cf. *tag-numbering*). Rappelons que pour chaque marquage tel qu'il apparaît dans le FEW, *tag-numbering* rétablit les marqueurs ellipsés au moyen d'un attribut *id* ajouté à la balise <pnum>. La première opération de l'algorithme *tag-microstructure* dépend donc de l'algorithme *tag-numbering* : elle consiste à chercher si le paragraphe contient une balise <pnum> et à en extraire les informations de marquage situées dans l'attribut *id*. Ces informations sont données telles quelles comme valeur à l'attribut *structid* de <p>.

Si le paragraphe traité ne commence pas par un marqueur alphanumérique, aucune balise <pnum> ne permet de déterminer sa place dans la structure de l'article. C'est la succession des paragraphes qui fournit alors un indicateur de reconnaissance :

> IS. Tout paragraphe de documentation sans marquage appartient nécessairement au même niveau structurel que le paragraphe précédent.

L'algorithme consiste à itérer sur chaque paragraphe de l'article dans l'ordre d'apparition dans le texte, en gardant en mémoire (dans une variable de type string) la valeur de l'attribut *structid* du paragraphe précédent et en mettant à jour cette variable en cas de détection d'un nouveau <pnum>. Si le premier paragraphe de l'article ne contient pas de balise <pnum>, la valeur de l'attribut *structid* pour ce paragraphe (et les paragraphes suivants éventuels sans <pnum>) est définie comme « none ».

```
|[
VAR
```

```
    prev_id : string ;
    ck : chunk ;
    p_licites : set of tags
BEGIN
prev_id := "unknown"
; p_licites := [<doc>, <mixt>]
; "initialiser ck à la première balise de l'article
   appartenant à p_licites"
; DO ck != null →
    "identifier le paragraphe"
    ...
    ; "déplacer ck à la prochaine balise appartenant à p_licites"
OD
END
]|
```

La commande `"identifier le paragraphe"` nécessite la déclaration d'une variable supplémentaire :

```
    ck_p : chunk

|[
"déplacer ck_p à la première balise ouvrante <pnum> après ck"
; IF ck_p != null →
    prev_id := "extraire de <pnum> la valeur de l'attribut id"
[] ck_p = null →
    SKIP
FI
; "ajouter à <p> un attribut structid prenant la valeur
   de la variable prev_id"
]|
```

5.4.24.3 Seconde phase : détection et balisage des groupements de lexèmes

> Dans chaque paragraphe, les lexèmes sont regroupés et hiérarchisés par la ponctuation. Des marqueurs textuels explicitent éventuellement les regroupements les plus importants.

Rappelons que seuls les groupements explicités par des marqueurs textuels sont reconnus et balisés dans le cadre de la rétroconversion (→ 3.7.2).

Les marqueurs textuels ont été identifiés et balisés par l'algorithme *tag-title*. La présence de balises <title> constitue donc une condition *sine qua non* pour identifier le début d'un groupement. Cependant, tous les éléments <title> ne sont pas à prendre en compte. L'algorithme *tag-title* a en effet balisé plusieurs sortes de marqueurs : les marqueurs connus (« ablt », « zuss » etc.), les marqueurs situés en début de paragraphe, les marqueurs entre crochets et les marqueurs internes. Parmi tous ces marqueurs, *tag-microstructure* ne conserve que ceux qui explicitent un groupement de lexèmes situé

au plus haut niveau de la hiérarchisation, c'est-à-dire les groupements signalés par un tiret (semi-)cadratin qui précède le marqueur textuel.

L'algorithme consiste dès lors à identifier les balises ouvrantes <title> dans le paragraphe et à vérifier la présence d'un tiret juste avant la balise. Un élément <title> précédé d'un tiret est dit valide. Lorsqu'un <title> valide est détecté, ce <title> et les lexèmes qui suivent jusqu'à la prochaine balise <title> valide (ou jusqu'à la fin du paragraphe) sont inclus dans un élément XML <group>. Cette opération est valable parce qu'on considère que les groupements marqués par un tiret et par un titre sont situés sur le même niveau (en d'autres termes, aucun <group> ne peut être inclus dans un autre <group>).

5.4.24.4 Génération d'un plan de l'article

La création d'un plan d'article consiste à itérer sur chaque paragraphe de documentation (paragraphes marqués <doc> et <mixt>) et à générer un court descriptif qui reprend le début du paragraphe, à savoir

1. le marquage alphanumérique (repris de l'attribut *structid*),
2. le marqueur textuel (<title>) débutant éventuellement le paragraphe,
3. la première cellule lexicale (<unit>) du paragraphe.

Les descriptifs de chaque paragraphe sont présentés dans l'ordre d'apparition de ces derniers. Le tout est fourni sous forme de commentaire XML. Voici un exemple de descriptif généré pour l'article *completus* (FEW 2 982b-983b) :

```
< !-- article map
I 1 a Vollständig. — Mfr. nfr. complet „à quoi il ne manque aucune des parties nécessaires“ (seit ca. 1300, Monstr ; Rhlitt 6, 464)
I 1 a Ablt. — Afr. mfr. completement „d'une manière complète“ (13. jh.—1503, TL ; Gdf ; RF 32, 83)
I 1 a Nfr. décompléter „rendre incomplet“ (seit 1779, Gohin)
I 1 a Nfr. incomplet „non complet“ (seit Ac 1762)
I 1 b Grundwort und ablt. in speziellen berufssprachlichen bed.
I 1 b α Nfr. complet adj. „(t. d'administr. milit.) (corps de troupes) qui atteint le nombre fixé de son effectif“ (seit Montaigne)
I 1 b β Nfr. complet „(t. de libr.) (livre) auquel il ne manque pas de feuilles, (ouvrage) auquel il ne manque pas de volumes“ (seit Widerh 1675)
I 1 b γ Nfr. complet „(habit) auquel il ne manque aucune des pièces nécessaires“ (seit mitte 17. jh.)
I 1 b δ Nfr. fleur complète „(t. de bot.) fleur qui a un calice, une corolle, une ou plusieurs étamines et un ou plusieurs pistils“ (seit Trév 1771)
I 2 Vollkommen. — Mfr. nfr. complet „(homme) qui a toutes les qualités désirables ; (joie, succès, etc.) qui ne laisse rien à désirer “ (seit 16. jh., doch bis 19. jh. selten, Sand)
I 2 Ablt. — Nfr. compléter (un homme) „le rendre parfait“ (seit Hugo 1835)
II Apr. completiu „qui sert à compléter“ (14. jh.)
III 1 Afr. complie „dernière partie de l'office qui se dit ou se chante après vêpres“ (12.—14. jh., Gdf ; TL ; Littré ; Brendan W 232 ; Mon Guill ; Pass)
III 2 Mfr. completoires „complies“ (16. jh.)
-->
```

5.4.24.5 Résumé

Dépendances

Tag-microstructure dépend de *tag-title*, *tag-numbering*, *split-doc-com*, *split-mixt-art*, *tag-unit* et *merge-mixt-p*.

Description

Tag-microstructure identifie et balise la microstructure d'un article, d'une part en ajoutant à chaque paragraphe <p> un attribut *structid* contenant la position du paragraphe dans la structure alphanumérique de l'article, d'autre part en identifiant par une balise <group> les groupements de lexème qui, au sein d'un paragraphe, sont explicités par un marqueur textuel.

Seuls les paragraphes <doc> et <mixt> sont traités par *tag-microstructure*. Le traitement d'un paragraphe s'effectue en trois phases : l'attribution au paragraphe d'un marquage alphanumérique, l'identification des marqueurs textuels pertinents et le balisage des groupements de lexèmes sur lesquels portent ces marqueurs.

L'attribution du marquage alphanumérique s'effectue en cherchant la valeur de l'attribut *id* contenu dans la balise <pnum>. En l'absence de <pnum> (signifiant probablement que le paragraphe n'est pas le premier du groupe auquel il appartient), le paragraphe reçoit le marquage donné au paragraphe précédent. Si le premier paragraphe de l'article ne contient pas de <pnum>, l'attribut *structid* reçoit une valeur conventionnelle.

L'identification des marqueurs textuels pertinents consiste à itérer sur les balises <title> et à les mémoriser au moyen d'une structure de données appropriée. Seuls les titres commençant le paragraphe ou précédés d'un tiret (semi-)cadratin sont conservés. Ce traitement s'effectue dans une chaîne virtuelle rendant invisibles la plupart des balises (cf. partition).

Le balisage des groupements de lexèmes s'effectue uniquement si un titre au moins a été détecté comme valide. Dans ce cas, tous les groupes commençant par un titre sont balisés <group>, depuis le titre en question (inclus dans le balisage) jusqu'au titre suivant (exclu du balisage) ou jusqu'à la fin du paragraphe. Les tirets (semi-)cadratins qui précèdent les titres ne sont pas inclus dans le balisage.

Tag-microstructure génère ensuite un plan de l'article, ajouté en tête d'article sous forme de commentaire XML, qui reprend le début de chaque paragraphe (*structid*, marqueur textuel éventuel et première cellule).

5.5 Algorithmes de post-traitement

La rétroconversion d'un article selon les algorithmes décrits précédemment conduit à plusieurs documents XML (un par algorithme) contenant le texte du FEW et, dans celui-ci, une série de balises identifiant les types d'information importants. Le document XML produit par le dernier algorithme de la séquence contient l'article du FEW finalement rétroconverti. Après ce dernier algorithme, aucune autre balise n'est insérée dans le texte de l'article.

Durant la rétroconverison, des erreurs de balisage ont pu se produire. Les avertissements produits dans certains contextes par les algorithmes de balisage constituent un premier moyen de détection de ces erreurs. Tous ces avertissements sont consignés à la fin du document XML, dans des balises <provenance> indiquant le résultat de chacun des algorithmes :

> <provenance op="tag-biblio" timestamp="1294908580486" type="algo">
> <provenance_update count="79" tag="biblio" type="spliced tags"/>
> <provenance_warning msg="warning : cannot decide whether keyword |Montaigne| should be tagged with <geoling> or <biblio>"/>

Ces avertissements ne sont toutefois pas suffisants. Il est possible que certaines informations soient passées au travers des mailles du filet, par exemple en cas de non-exhaustivité des listes de mots-clés utilisées. Une série d'algorithmes, dits de post-traitement puisqu'ils sont appliqués au document XML rétroconverti, ont pour objectif de remédier aux limitations des algorithmes de balisage en permettant la détection de ce type d'erreurs. Il s'agit des algorithmes *show-untagged-text*, *show-untagged-unit-text*, *show-tags* et *show-isolated-tags*.

Enfin, un algorithme, *export-units-to-xs*, a pour objectif d'extraire de l'article rétroconverti chacune des cellules lexicales (balisées lors de la rétroconversion au moyen de l'élément XML <unit>) et d'y associer toutes les informations qui les concernent directement, à l'exclusion des informations liées à la structure de l'article. Le résultat fourni par cet algorithme constitue, en quelque sorte, l'explicitation de la dimension thesaurus du FEW. La liste des cellules, présentée en fin du document sous forme de commentaire XML, est exportable telle quelle dans le logiciel XS utilisé par les rédacteurs de la refonte, ce qui doit permettre une interaction entre les deux projets d'informatisation.

5.5.1 Identification des parties non balisées : show-untagged-text

L'algorithme *show-untagged-text* a pour objectif de mettre en évidence les segments de texte qui n'ont pas été attribués à un type d'information précis par les algorithmes de balisage. Il repère notamment les extraits de texte qui, dans la documentation, n'ont pas été inclus dans une cellule lexicale (<unit>).

Pour chaque paragraphe de l'article traité, y compris l'entrée et les notes, *show-untagged-text* détermine un taux de balisage, exprimé en pourcentage. Tous les caractères du FEW non triviaux (c'est-à-dire hors espacements et ponctuation) qui se trouvent uniquement dans des balises typographiques (<p>, <b>, <e>, <i>, <sc>) et/ou de paragraphe (<entry>, <doc>, <mixt> etc.) sont considérés comme non balisés. Le pourcentage donné représente le nombre de caractères balisés par rapport au nombre total de caractères du paragraphe. Par exemple, si le nombre de caractères non balisés est de 2du paragraphe est 98

Le taux de balisage de chaque paragraphe, ainsi que les extraits de texte non balisés, sont donnés sous la forme d'un commentaire XML qui apparaît à la fin de l'article du

FEW, comme suit (FEW 3, 451b-452a, FEMININUS) :

```
< !-- show-untagged-text
Untagged text in entry (tagging rate : 47%)
<...> weiblich.
Untagged text in <mixt> paragraph (tagging rate : 89%)
<...>Daraus entlehnt <...> ; verkürzt <...>. Mit suffw. <...>
Untagged text in <doc> paragraph (tagging rate : 100%)
<...>
Untagged text in <notes> paragraph (tagging rate : 39%)
<...>1) In dieser bed. auch substantiviert zur über-<...>setzung des Goethewortes
<...>
Untagged text in <notes> paragraph (tagging rate : 11%)
<...>2) Nach dem <...>femineus<...>. So bei Baude-<...>laire, doch heute veral-
tet.<...>
-->
```

Ce commentaire permet, en un coup d'œil,

- d'estimer rapidement le résultat de la rétroconversion pour l'article concerné ;
- de détecter, parmi les extraits de texte non balisés, les sigles bibliographiques et géolinguistiques qui n'ont pas été reconnus par les algorithmes en raison de la non-exhaustivité des listes de mots-clés ;
- de repérer les incohérences du FEW (sigles non balisés car non conformes) et de les soumettre à des experts pour correction éventuelle.

5.5.2 Identification des parties de cellules lexicales non balisées : show-untagged-unit-text

L'algorithme *show-untagged-unit-text* complète l'algorithme show-unit-text en mettant en évidence les extraits de texte qui, à l'intérieur d'une cellule lexicale (<unit>), n'ont pas été balisés. La liste de ces extraits de texte est fournie en fin d'article, après celle de *show-untagged-text*, de la même façon que cette dernière.

Soulignons que cette liste et, de manière générale, toutes les listes générées par les algorithmes de post-traitement apparaissent uniquement si les algorithmes de post-traitement sont activés : il est tout à fait possible de désactiver l'un ou l'autre algorithme de post-traitement si désiré, au moyen du fichier *few-config.txt* qui est fourni par le logiciel de rétroconversion.

5.5.3 Identification des parties balisées : show-tags

L'algorithme *show-tags* constitue le pendant de *show-untagged-text*, puisqu'il a pour objectif de fournir une liste de toutes les sections de texte fewien qui ont été balisées.

Les balises et leur contenu sont présentés pour chaque paragraphe de l'article (en-

trée et notes inclus), de la façon suivante :

```
< !-- show-tags
Tags inserted into entry, column 94b
<b>accusativus</b>
<etymon>accusativus</etymon>
Tags inserted into <doc> paragraph, column 94b
<unit>Mfr. nfr. accusatif m. „cas auquel on met le complément direct“ (seit ca.
1170, EdConf, FrMod 21, 217)</unit>
<geoling>Mfr.</geoling>
<geoling>nfr.</geoling>
<form>accusatif</form>
<i>accusatif</i>
<gram>m.</gram>
<def>„cas auquel on met le complément direct“</def>
<precisions>(seit ca. 1170, EdConf, FrMod 21, 217)</precisions>
<attestation>seit ca. 1170, EdConf, FrMod 21, 217</attestation>
<date>ca. 1170</date>
<biblio>EdConf</biblio>
<biblio>FrMod 21, 217</biblio>
<unit>adj. „qui concerne l’accusatif“ (1380, Aalma 98 ; Pom 1671–1700 ; Lar
1866–1948)</unit>
<gram>adj.</gram>
<def>„qui concerne l’accusatif“</def>
<precisions>(1380, Aalma 98 ; Pom 1671–1700 ; Lar 1866–1948)</precisions>
<attestation>1380, Aalma 98</attestation>
<date>1380</date>
<attestation> Pom 1671–1700</attestation>
<biblio>Pom 1671</biblio>
<date>1700</date>
<attestation> Lar 1866–1948</attestation>
<biblio>Lar 1866–1948</biblio>
Tags inserted into <com> paragraph, column 94b
<lang>lt.</lang>
<i>accusativus</i>
<etymon>accusativus</etymon>
-->
```

Cette présentation, insérée en fin d’article en tant que commentaire XML, doit permettre de vérifier, de façon plus rapide et plus confortable que ne le serait la lecture du document XML, la justesse du balisage inséré par les algorithmes de rétroconversion.

5.5.4 Identification des balises occupant une place suspecte : show-isolated-tags

L’algorithme *show-isolated-tags* complète show-tags en mettant en évidence

- d’une part, les éléments XML correspondant à des molécules du FEW (extraits de texte balisés <biblio>, <date>, <def>, <form>, <geoling> et <gram>) qui se trouveraient en dehors d’une cellule <unit> ;
- d’autre part, les éléments XML ne correspondant pas à des molécules du FEW qui se trouveraient à l’intérieur d’une cellule <unit>.

La liste de ces éléments, présentée en fin d'article sous forme de commentaire XML, doit permettre de détecter rapidement des situations potentiellement anormales, qui pourraient être un indice d'erreurs de balisage.

5.5.5 Extraction des cellules lexicales : export-units-to-xs

L'algorithme *export-units-to-xs* n'est pas à proprement parler un algorithme de post-traitement, puisqu'il est appelé par l'algorithme de balisage *tag-unit*. Néanmoins, nous le décrivons ici, car son résultat apparaît en fin de document de la même façon que le résultat des algorithmes de post-traitement.

Cet algorithme a pour objectif d'extraire, hors de l'article rétroconverti, une liste de toutes les cellules lexicales. Cette liste reprend pour chaque cellule toutes les molécules, y compris les molécules ellipsées. Chaque cellule est présentée individuellement, avec reprise de son adresse FEW et de l'étymon. Dans le cas où la cellule fait partie d'un groupement identifié comme tel par un marqueur textuel (cf. *tag-microstructure*), ce marqueur est également repris. La liste résultante est présentée comme suit :

< !-- export-units-to-xs

<fiche etymon="accusativus" lang="Mfr." lang="nfr." forme="accusatif" gram="m." def="„cas auquel on met le complément direct“" ref="(seit ca. 1170, EdConf, FrMod 21, 217)" N="FEW 24/1, 94b, ici 1, §1, u1"></fiche>

<fiche etymon="accusativus" lang="(imp.) Mfr." forme="(imp.) accusatif" gram="adj." def="„qui concerne l'accusatif“" ref="(1380, Aalma 98 ; Pom 1671–1700 ; Lar 1866–1948)" N="FEW 24/1, 94b, ici 1, §1, u2"></fiche>

-->

Export-units-to-xs est intitulé comme tel car il fait le lien avec le logiciel XS utilisé par la rédaction du FEW pour la refonte des articles de la tranche alphabétique B- (cf. Matthey et Nissille 2010). Le format de présentation de la liste est en effet celui qui est attendu par le logiciel XS. L'exportation dans XS de la liste générée doit permettre que le contenu des articles rétroconvertis puisse être traité par ce logiciel de la même façon que celui des articles rédigés au fur et à mesure. Il est important de souligner ici que l'exportation vers XS concerne uniquement la liste des cellules, donc la dimension thesaurus du FEW. La dimension monographique (structure de l'article et analyse linguistique qui en découle) n'est pas conservée dans le processus d'exportation.

5.6 Séquençage des algorithmes

5.6.1 Dépendances entre algorithmes

Les algorithmes de balisage décrits ci-dessus comportent des dépendances : leur application est conditionnée par l'application préalable d'autres algorithmes. Plusieurs algorithmes de balisage se justifient par leur intérêt indirect (c'est-à-dire par rapport à un autre algorithme) autant que par leur intérêt direct (c'est-à-dire par rapport à la modélisation du FEW). Le balisage des définitions (*tag-def*) et des catégories grammaticales (*tag-gram*) présente par exemple peu d'intérêt pour l'exploitation même du

FEW, mais il est essentiel pour permettre la reconnaissance des cellules lexicales par *tag-unit*.

L'ordre d'application des algorithmes joue donc un rôle crucial dans le succès de la rétroconversion. L'impossibilité de trouver un séquençage des algorithmes qui respecte toutes les dépendances remettrait en question les algorithmes eux-mêmes, en demandant de trouver d'autres critères que ceux qui ont été définis, de façon à supprimer certaines dépendances problématiques. Les dépendances cycliques (un algorithme A dépendant d'un algorithme B, qui dépendrait lui-même de A) sont notamment interdites.

5.6.2 Graphe des dépendances

Les dépendances entre algorithmes peuvent être représentées visuellement par un graphe orienté. Un nœud du graphe représente un algorithme, tandis qu'un arc représente une dépendance. Un arc est orienté de façon à signifier « doit être appliqué avant ».

Dans la rétroconversion du FEW, la multitude des relations entre algorithmes rend le graphe peu lisible si nous le représentons en un seul bloc. Aussi le graphe des dépendances est-il exposé ci-dessous en cinq parties successives.

L'examen de ces dépendances doit nous permettre de définir un séquençage d'algorithmes, c'est-à-dire un chemin qui respecte l'orientation des arcs. Il est important de remarquer qu'un arc entre deux algorithmes n'oblige pas ces deux algorithmes à se succéder directement dans la séquence : le fait qu'un algorithme A doive être appliqué avant un algorithme B n'empêche pas d'intercaler entre A et B un autre algorithme C. Par ailleurs, il est possible de passer d'un algorithme à l'autre même si aucune dépendance ne les relie. La seule contrainte est que le chemin choisi ne peut aller en sens inverse d'un arc. Un chemin entre plusieurs algorithmes est dit valide s'il respecte le sens de tous les arcs.

5.6.2.1 Algorithmes de prétraitement

Les huit algorithmes de prétraitement comportent entre eux sept dépendances, qui peuvent être schématisées comme à la figure 5.12.

Les dépendances permettent plusieurs chemins possibles parmi ces algorithmes, de gauche à droite et de haut en bas.

5.6.2.2 Algorithmes de balisage (1)

Parmi les sept algorithmes de balisage présents dans le sous-graphe de la figure 5.13, seuls *tag-entry* et *tag-notes* sont complètement indépendants (les dépendances aux algorithmes de prétraitement étant mises à part). Les cinq autres algorithmes dépendent, directement ou indirectement, de l'un de ces deux algorithmes. *Tag-entry* et *tag-notes* sont donc à appliquer au tout début de la séquence, après les algorithmes de prétraitement. En revanche, le graphe montre qu'aucun algorithme ne dépend de *tag-affix*, de *tag-lang-etymon* et de *tag-concept*. Les cylindres représentent les dépendances à des listes de mots-clés, le nombre 2 indiquant que ces listes sont au nombre de deux (en allemand et en français).

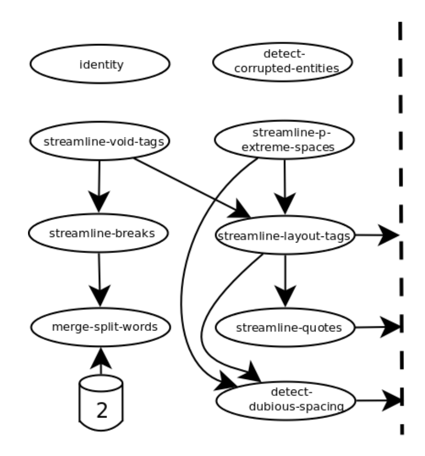


FIGURE 5.12 – Algorithmes de prétraitement

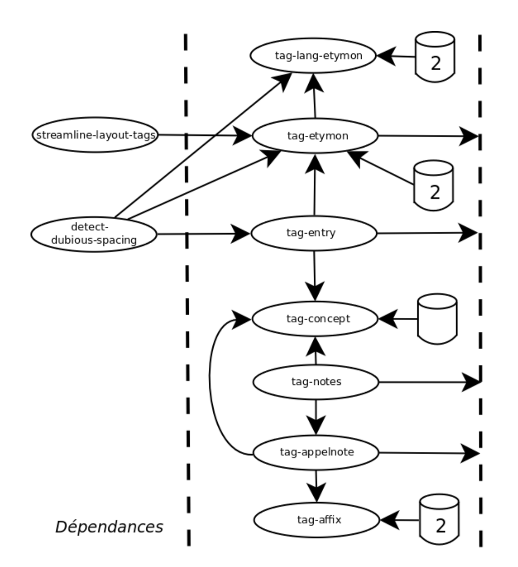


FIGURE 5.13 – Algorithmes de balisage (1)

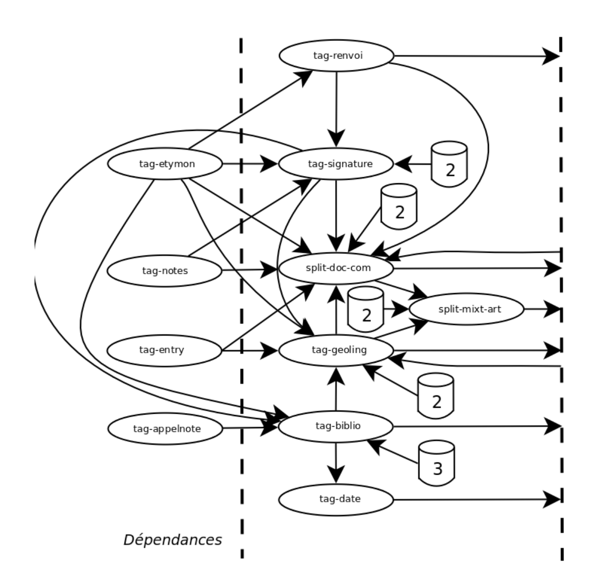


FIGURE 5.14 – Algorithmes de balisage (2)

5.6.2.3 Algorithmes de balisage (2)

Les sept algorithmes de balisage présentés dans la figure 5.14 dépendent, directement (pour six d'entre eux) ou indirectement (pour *tag-date*) de quatre algorithmes présentés dans le sous-graphe précédent, à savoir *tag-entry*, *tag-notes*, *tag-appelnote* et *tag-etymon*. L'algorithme *split-doc-com* est particulièrement dépendant, puisqu'il nécessite l'application préalable de sept autres algorithmes.

5.6.2.4 Algorithmes de balisage (3)

Les quatre algorithmes exposés dans la figure 5.15 dépendent, directement ou indirectement, des algorithmes précédents. L'algorithme *tag-form* est particulièrement dépendant (cinq dépendances, parmi lesquelles les trois autres algorithmes). L'algorithme *tag-def* joue un rôle important, puisqu'il conditionne directement le succès de quatre algorithmes, dont deux du groupe précedent.

5.6.2.5 Algorithmes de balisage (4)

L'algorithme *tag-attestation* mis à part, les algorithmes présentés dans le sous-graphe de la figure 5.16 dépendent d'un grand nombre d'autres algorithmes. Cette situation s'explique par le fait que ce sont majoritairement des algorithmes de groupement, qui imposent une structure aux types d'information du FEW. La reconnaissance des cellules lexicales (*tag-unit*), notamment, ne pourrait réussir sans la détection préalable de toutes les molécules qui composent la cellule.

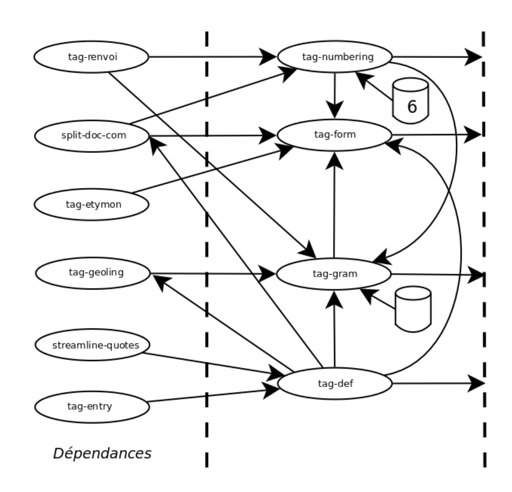


FIGURE 5.15 – Algorithmes de balisage (3)

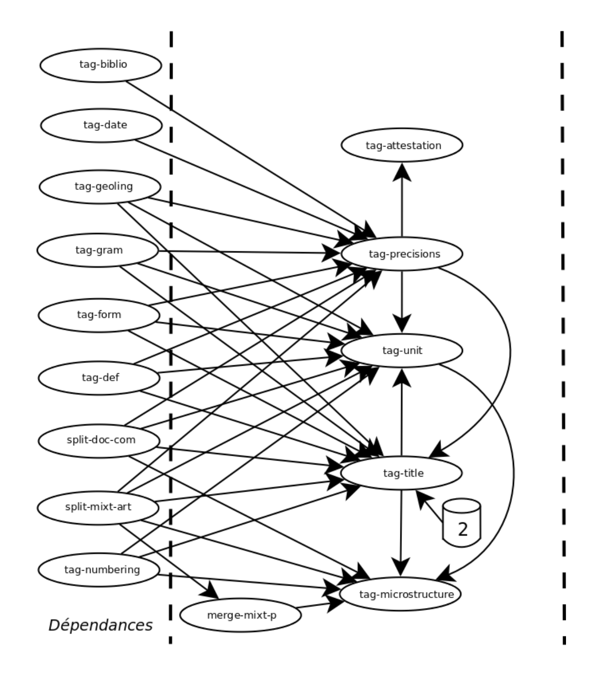


FIGURE 5.16 – Algorithmes de balisage (4)

5.6.3 Choix du séquençage

En fonction des arcs définis ci-dessus, plusieurs chemins sont possibles. Ci-dessous se trouve le séquençage des algorithmes (hors algorithmes de post-traitement) qui a été choisi après analyse des possibilités. Ce séquençage respecte l'orientation de tous les arcs.

1. detect-corrupted-entities
2. streamline-p-extreme-spaces
3. streamline-void-tags
4. streamline-breaks
5. merge-split-words
6. streamline-layout-tags
7. streamline-quotes
8. detect-dubious-spacing
9. tag-notes
10. tag-appelnote
11. tag-entry
12. tag-etymon
13. tag-lang-etymon
14. tag-renvoi
15. tag-signature
16. tag-def
17. tag-biblio
18. tag-geoling
19. split-doc-com
20. tag-numbering
21. split-mixt-art
22. tag-affix
23. tag-date
24. tag-gram
25. tag-form
26. tag-concept
27. tag-precisions
28. tag-attestation
29. tag-title
30. tag-unit
31. merge-mixt-p
32. tag-microstructure

Ce séquençage est valable dans le cadre d'une saisie manuelle du FEW respectant le format FFML défini comme format d'entrée du logiciel de rétroconversion. S'il s'avérait que le FEW ne pouvait bénéficier d'une saisie manuelle et que le format FFML ne pouvait être garanti, la rétroconversion deviendrait plus compliquée. D'une part, le séquençage des algorithmes devrait être revu. L'absence de balisage préalable de l'article (<art>) et de l'entrée (<h>) rendrait en effet la reconnaissance de l'étymon-vedette essentielle pour la reconnaissance de l'article et de l'entrée. *Tag-etymon* se trouverait dès lors déplacé avant *tag-entry*. D'autre part, quelques-uns des algorithmes devraient être revus, parmi lesquels tag-notes (qui ne bénéficierait plus du balisage préalable du champ des notes, → 5.4.1) et *tag-entry*.

5.7 Conclusion

Les algorithmes constituent le noyau du logiciel de rétroconversion : dès lors, ils déterminent la question de la faisabilité de l'opération toute entière. Heureusement, les conclusions de l'étude menée dans ce chapitre sont largement positives. En effet, un algorithme de détection a pu être mis au point pour chacun des types d'information prévus dans la modélisation du FEW. La complexité structurelle du FEW n'empêche pas de trouver des combinaisons d'indicateurs assez fiables et spécifiques pour reconnaître chaque type d'information de façon satisfaisante, en évitant les problèmes de circularité. La traduction des indicateurs dans une représentation informatique, l'utilisation d'outils de détection de motifs (listes de mots-clés et expressions régulières) et le mécanisme des chaînes virtuelles (qui pallie aux problèmes de collisions entre mot-clés et de balises parasites dans les expressions régulières) permettent l'automatisation du processus de reconnaissance. Les 37 algorithmes élaborés sont applicables dans un ordre qui respecte les dépendances, de sorte que les types d'information dont la reconnaissance est nécessaire à la détection d'autres types d'information sont balisés préalablement à ces derniers. En théorie, la faisabilité de la rétroconversion est bel et bien acquise.

Cette conclusion positive est évidemment à nuancer : pour obtenir ce résultat, il a fallu accepter quelques imperfections. L'élaboration d'algorithmes parfaits, détectant infailliblement chaque occurrence de chaque type d'information du FEW, demanderait de maîtriser l'ensemble des incohérences de l'ouvrage, ce qui est impossible. Le FEW ne s'automatise pas : il s'apprivoise. L'utilisation de listes de mots-clés est un exemple d'approche imparfaite sans laquelle la rétroconversion n'eût pas été possible. Une liste constitue en effet un moyen de détection très fiable, à condition d'être exhaustive et de correspondre exactement à la représentation informatique du texte. Or, il est impossible actuellement, même aux rédacteurs du FEW, de répertorier sans oubli l'ensemble des données de l'ouvrage. Nous avons pourtant utilisé un grand nombre de ces listes, tout en sachant que leur exhaustivité n'était pas vérifiable pour le FEW dans l'état actuel de son informatisation. Un certain nombre d'informations sont donc susceptibles de ne pas être reconnues par les algorithmes de balisage.

Afin de remédier à ce problème et de générer des procédures d'autocorrections peu coûteuses (complétage des listes), quelques garde-fous ont été mis en œuvre, parmi lesquels les algorithmes de post-traitement, qui doivent faciliter la détection des erreurs et des oublis. Une autre conséquence de l'imperfection des algorithmes est la modification – légère – du modèle XML primitivement défini lors de la modélisation. Le schéma

XML résultant de la rétroconversion (→ G.2) est plus souple et moins précis que celui du modèle. Dans les précisions par exemple, le modèle prévoit des <punit>, mais le résultat de la rétroconversion ne les identifie pas. Seul un attribut *status*="contains_unit" donne une indication sur la possibilité que des cellules lexicales apparaissent dans les précisions.

Les adaptations apportées au schéma initial ont trois objectifs. Tout d'abord, rendre ce dernier plus flexible et plus permissif par rapport aux incohérences du FEW et aux erreurs de balisage. La possibilité d'éléments <gram> contenant un marqueur de renvoi bibliographique (cf. *tag-precisions*) est un exemple d'adaptation destinée à permettre de façon temporaire des errreurs de balisage, donc des faux positifs. Sans cette possibilité de balisage erroné, le balisage des catégories grammaticales conduirait massivement à des faux négatifs, ce que nous avons cherché à éviter. Un deuxième objectif du schéma final est de signaler les cas ambigus, requérant vérification par un expert. Cet objectif a conduit à l'ajout, dans plusieurs éléments XML, d'attributs qui n'ont aucun intérêt pour l'exploitation du FEW, mais qui sont utiles à la vérification du résultat de la rétroconversion. L'attribut *status* assigné aux appels de note (→ 5.4.2) en est un exemple. Enfin, certaines adaptations du schéma initial sont d'ordre technique : elles doivent faciliter le traitement informatique. C'est le cas des modifications concernant l'adresse FEW de l'article (répartie sur les deux balises <art> et <col> au lieu de la seule balise <art>, → 3.5.1) ou concernant le balisage <note> inclus dans le paragraphe <p> (alors que les balises <doc> et <com> entourent le paragraphe). Ces adaptations du schéma peuvent surprendre un linguiste par leur incohérence apparente, mais elles permettent d'optimiser les opérations de recherche de balises lors de l'implémentation des algorithmes.

Il nous semble que ces quelques modifications, qui représentent autant d'imperfections formelles, par rapport au modèle initial ne portent pas atteinte à l'essentiel : le balisage que les algorithmes insèrent dans un article du FEW devrait permettre la plupart des requêtes souhaitées par les utilisateurs. L'implémentation des algorithmes, la rétroconversion effective d'un corpus représentatif du FEW et l'analyse du résultat, proposée dans le chapitre suivant, constituent une dernière étape pour valider en pratique l'étude théorique que nous venons d'exposer.

Chapitre 6

Analyse et exploitation des résultats

6.1 Introduction

Les algorithmes de rétroconversion du FEW qui ont été décrits dans le chapitre précédent ont été implémentés en Java par Cyril Briquet (ATILF) entre octobre 2008 et décembre 2009. Leur implémentation et celle de tous les composants annexes sous-jacents nécessaires à leur fonctionnement (dont le mécanisme de construction des chaînes virtuelles, → 4.4.2) a nécessité le développement d'un logiciel complet qui compte actuellement environ 75 000 lignes de code.

Après finalisation de ce logiciel, il a été possible de tester les algorithmes sur corpus. Deux corpus de test ont été constitués dans ce but.

Le premier est un corpus d'analyse, sur la base duquel les algorithmes ont été revus jusqu'à ce que leur application sur le corpus fonctionne avec un résultat jugé satisfaisant. Ce premier corpus se compose en partie de 52 articles du projet ANR DETCOL(Développement et Exploitation Textuelle d'un Corpus d'Oeuvres Linguistiques [1]), constituant un tout linguistique cohérent et apparaissant dans les volumes 2, 3, 4, 6, 7, 8, 9, 10, 12, 13, 14 et 24 du FEW. À cet ensemble ont été ajoutés 78 articles des volumes 3, 6, 16, 19, 21 et 25 du FEW, choisis de façon à compléter l'éventail des types d'information que présentent les articles DETCOL. L'ensemble de ces 130 articles couvre la plupart des volumes du FEW, y compris ceux consacrés aux matériaux d'origine inconnue et ceux de la refonte de la tranche alphabétique *A-*.

Le second corpus de test, beaucoup plus modeste, a été constitué dans le but de vérifier la validité, sur de nouveaux articles, des algorithmes construits sur la base du premier corpus. Il se compose en partie de 10 articles supplémentaires du projet DETCOL et en partie de 5 autres articles choisis à travers l'ensemble du FEW. À ces 15 articles a été ajouté, à la demande des rédacteurs du FEW, un article de la refonte de la tranche alphabétique B-.

[1] Pour une présentation du projet : `http://ctlf.ens-lsh.fr/articles/documents/ct_projet_detcol.pdf`.

Dans ce chapitre, nous nous proposons d'examiner le résultat de la rétroconversion de ces deux corpus. Dans un premier temps (→ 6.2), nous analyserons un seul article de façon approfondie, afin de voir si le balisage inséré par les algorithmes reflète de façon satisfaisante la modélisation proposée. Nous évaluerons ensuite le balisage inséré dans le premier corpus, afin de donner une idée des types d'information qui sont bien ou moins bien reconnus sur l'ensemble des 130 articles (→ 6.3). Nous ferons de même avec le second corpus (→ 6.4). Enfin, nous émettrons des propositions quant aux modalités d'exploitation de ces articles rétroconvertis (→ 6.5).

6.2 Exemple d'article rétroconverti

L'article CHOCOLATL (FEW 20, 63b-64a) nous servira d'exemple. Nous l'avons choisi, non pas en raison du référent de l'étymon, mais parce que sa taille, sa structure et son contenu en font un bon exemple d'article type du FEW.

La version FFML de l'article a été obtenue après numérisation et correction manuelle. Nous avons ensuite soumis cette version au logiciel de rétroconversion. Le document XML final produit par le logiciel est fourni en annexe (→ G).

6.2.1 Explicitation de la dimension monographique

L'informatisation du FEW est censée, conformément à la modélisation proposée, faciliter l'utilisation de l'ouvrage dans ses deux dimensions : en tant que recueil de monographies et en tant que thesaurus (→ 2.7). Le balisage de l'article CHOCOLATL remplit-il le premier objectif, à savoir faciliter la lecture de l'article en tant qu'ensemble structuré, en explicitant sa dimension monographique ?

Les attributs des deux balises <art> et <col> de l'article rétroconverti en situent le début à l'intérieur du FEW (volume="20", pg="63", s="b"), indiquent qu'il suit le schéma général d'une documentation suivie d'un commentaire (type="doc-com") et signalent la métalangue (lang="german") :

<art book="1" ici="2" id="0" lang="german" type="doc-com" volume="20">
<col merge-split-word="no" pg="63" s="b"/>

L'auteur de l'article, non cité explicitement dans la version imprimée, a été identifié comme étant von Wartburg (<signature author="Wartburg" lang="german"/>).

Le document rétroconverti commence par un commentaire XML qui fournit une vision synthétique de la structuration de l'article :

< !– article map
1 Mfr. chocholate m. „breuvage fait avec des amandes de cacao“ (1598)
1 Ablt. — Nfr. chocolatière f. „vase où l'on prépare, où l'on sert le chocolat en boisson“ (seit 1680)
2 Nfr. chicolate f. „chocolat“ (1658)
–>

D'un coup d'oeil, nous constatons que la partie documentaire de l'article CHOCOLATL se compose de trois paragraphes, structurés en deux parties numérotées, et que le second paragraphe de la première partie reprend des dérivés (*Ablt.*). Ce plan synthétique n'indique pas que le premier paragraphe contient également une section

reprenant des lexèmes présentant un sens figuré (*Übertragen*). Toutefois, cette section a été correctement identifiée et balisée (<group>...</group>) dans le texte :

> —<group><title> Übertragen </title> Paris chocolat „personne, animal, objet de couleur chocolat“ [...]. </group>

Les références renvoyant, dans le commentaire, au marquage alphanumérique de la documentation (1 et 2) ont été identifiées (<pref>), ce qui permettra la création de liens hypertextes entre les marqueurs de la documentation et leur explicitation :

> <pref id="1" status="ok">1</pref> ist aus sp. chocolate entlehnt [...]. Unklar ist auch das verhältnis von <pref id="2" status="ok">2</pref> zu <pref id="1" status="ok">1</pref>. [...] Der erste beleg von <pref id="2" status="ok">2</pref> kommt von den kleinen Antillen [...]

L'appel de note présent à la fin du premier paragraphe a lui aussi été détecté et a reçu un identifiant qui l'associe à la note correspondante, ce qui permettra au lecteur de lire la note sans devoir se rendre à la fin de l'article :

> [...] „battu au jeu ; n'ayant pas réussi“ B <appelnote id="1" status="ok">
> <e>1)</e>
> </appelnote>. [...]
> <notes call-sequence="ok" count="1"> <p><note id="1">1) Diese bed. ist eine synonymalablt. [...]

Les renvois internes à l'article, entre la documentation, le commentaire et les notes, ont donc été reconnus. Aucun renvoi à un autre article du FEW n'a été détecté et, en effet, l'article n'en contient pas. En ce qui concerne les renvois externes au FEW, 22 sigles bibliographiques ont été balisés :

- dans le premier paragraphe : Trév 1732 ; Rich 1680 ; Trév 1732 ; SavBr 1723 ; A ; B ; B ;
- dans le deuxième paragraphe : Rich 1680 ; Trév 1771, 2, 148 ;
- dans le troisième paragraphe : A ;
- dans le commentaire : Doch ; MPh 23, 325 ; KZ 61, 93 ; König 66 ; Friederici 182 ; Corom 2, 75 ; Wid 1669 ; SavBr 1679, 2, 176 ; Der ; Friederici ;
- dans les notes : S ; DauzatArgot 80.

Parmi ces 22 éléments identifiés comme sigles bibliographiques se trouvent trois erreurs, situées dans le commentaire et les notes (*Der* est ici un article allemand, *Doch* une conjonction, et *S* un marqueur de renvoi, les majuscules étant dues au fait que ces items se situent en début de phrase). Un sigle, *Wid,*, a été marqué comme ambigu, parce qu'il appartient à la fois à la liste des sigles bibliographiques et à celle des étiquettes géolinguistiques (→ 5.4.9). Cette ambiguïté est signalée par l'attribut *status* de la balise <biblio> et est notifiée explicitement à la fin du document par un avertissement :

> <biblio status="geoling ?">Wid 1669</biblio> berichtet als [...]

> <provenance op="tag-biblio" timestamp="1295256561517" type="algo">
> <provenance_update count="22" tag="biblio" type="spliced tags"/>
> <provenance_warning msg="warning : cannot decide whether keyword |Wid| should be tagged with <geoling> or <biblio>"/>

Cet avertissement permet de détecter et de corriger l'erreur de balisage éventuelle,

ce qui n'est pas nécessaire ici, puisqu'il s'agit bien d'un sigle bibliographique.

Le balisage des sigles bibliographiques permettra un lien hypertextuel avec leur explicitation fournie par le *Beiheft* (→ 6.5.1). Il en va de même pour le balisage des étiquettes géolinguistiques et des mentions de langue non galloromanes, qui apparaissent respectivement 21 et 6 fois dans l'article, comme l'indique la balise <provenance> associée à l'algorithme *tag-geoling* :

```
<provenance op="tag-geoling" timestamp="1295378950829" type="algo">
<provenance_update count="21" tag="geoling" type="spliced tags"/>
<provenance_update count="6" tag="lang" type="spliced tags"/>
</provenance>
```

Les six mentions de langues qui ont été balisées se trouvent dans le commentaire, comme attendu (*sp.*, *sp.*, *it.*, *d.*, *piem.*, *Poschiavo*). Le sens de chacune de ces indications est précisée dans le *Beiheft*.

En résumé, le balisage inséré de façon automatique par le logiciel dans l'article CHOCOLATL est conforme à nos attentes en ce qui concerne l'explicitation de la dimension monographique : il situe l'article dans le FEW, il résume sa structure et il permet la création de liens hypertextuels pour les informations explicitées ailleurs, que ce soit dans une autre partie de l'article (appels de note, marquage alphanumérique) ou dans le *Beiheft* (sigles bibliographiques et géolinguistiques).

6.2.2 Explicitation de la dimension thesaurus

Le balisage de l'article CHOCOLATL remplit-il le second de nos objectifs, à savoir faciliter la consultation du FEW dans sa dimension thesaurus (→ 2.4.1) ?

En ce qui concerne les informations associées aux étymons, l'étymon-vedette et la langue de l'étymon-vedette ont été reconnus et balisés (<etymon lang="aztek." type="vedette">chocolatl</etymon>). Le type de transmission de l'étymon-vedette a également été reconnu, grâce à la mention de l'étymon dans le commentaire (<etymon descendance="emprunts" type="vedette">chocolatl</etymon>). Une consultation transversale du FEW via la langue d'origine des lexèmes ou via leur type de descendance (emprunts ou lexèmes héréditaires, le cas des créations internes n'étant pas traité) est donc rendue possible.

En ce qui concerne la consultation du FEW via les lexèmes, la balise <provenance> associée à l'algorithme *tag-unit* indique que 26 cellules lexicales ont été reconnues et balisées. Ces 26 cellules sont listées en fin de document (sous la forme voulue par le logiciel de rédaction modulaire de la refonte, → 5.5.5), avec rétablissement pour

chacune de l’étymon et de l’adresse FEW :

< !– export-units-to-xs
<fiche etymon="chocolatl" lang="Mfr." forme="chocholate" gram="m." def="„breuvage fait avec des amandes de cacao“" ref="(1598)" N="FEW 20/1, 63b, ici 2, §1, u1"></fiche>
<fiche etymon="chocolatl" lang="(imp.) Mfr." forme="chocolate" gram="(imp.) m." def="(imp.) „breuvage fait avec des amandes de cacao“" ref="(1640—Trév 1732)" N="FEW 20/1, 63b, ici 2, §1, u2"></fiche>
<fiche etymon="chocolatl" lang="(imp.) Mfr." forme="(imp.) chocolate" gram="f." def="(imp.) „breuvage fait avec des amandes de cacao“" ref="(1640—1672)" N="FEW 20/1, 63b, ici 2, §1, u3"></fiche>
<fiche etymon="chocolatl" lang="(imp.) Mfr." forme="chocolat" gram="m." def="(imp.) „breuvage fait avec des amandes de cacao“" ref="(seit 1666, Arveiller 178)" N="FEW 20/1, 63b, ici 2, §1, u4"></fiche>
<fiche etymon="chocolatl" lang="(imp.) Mfr." forme="chocolate" gram="(imp.) m." def="„substance solide de ce breuvage“" ref="(Rich 1680–Trév 1732)" N="FEW 20/1, 63b, ici 2, §1, u5"></fiche>
<fiche etymon="chocolatl" lang="(imp.) Mfr." forme="chocolat" gram="(imp.) m." def="(imp.) „substance solide de ce breuvage“" ref="(seit SavBr 1723)" N="FEW 20/1, 63b, ici 2, §1, u6"></fiche>
<fiche etymon="chocolatl" lang="lütt." forme="choûcolå" gram="(imp.) m." def= "„chocolat“" N="FEW 20/1, 63b, ici 2, §1, u7"></fiche>
<fiche etymon="chocolatl" lang="loch." forme="choucolâ" N="FEW 20/1, 63b, ici 2, §1, u8"></fiche>
<fiche etymon="chocolatl" lang="sav." forme="chôcolâ" N="FEW 20/1, 63b, ici 2, §1, u9"></fiche>
<fiche etymon="chocolatl" lang="mars." forme="chocoulat" ref="A" N="FEW 20/1, 63b, ici 2, §1, u10"></fiche>
<fiche etymon="chocolatl" lang="bearn" forme="chocolat" N="FEW 20/1, 63b, ici 2, §1, u11"></fiche>
<fiche etymon="chocolatl" lang="Saint." lang="SeudreS." forme="chacolat" N="FEW 20/1, 63b, ici 2, §1, u12"></fiche>
[...]–>

Dans chacune de ces cellules, les molécules ont été reconnues et balisées également. 18 balises <imp/> ont été insérées, qui rétablissent l’implicite dû à la dimension monographique du FEW (→ 2.6.3.1, 3.6.3.5). Ces explicitations concernent les quatre molécules obligatoires (étiquette géolinguistique, signifiant, catégorie grammaticale et signifié).

En ce qui concerne les précisions, l’algorithme *tag-unit* nous avertit qu’elles manquent dans 12 cellules lexicales, telle que la suivante :

<provenance_warning msg="warning : no <precisions> found in unit
| <geoling>lütt.</geoling> <form>choûcolå</form> <imp type=’gram’ contents=’m.’> <def>„chocolat“</def>|"/>

Après vérification (dans l’article même, ou via la liste des cellules affichée en fin de document par les algorithmes de post-traitement (*show-tags* et *export-units-to-xs*), il s’avère que l’absence de précisions pour ces cellules est conforme au texte du FEW et que le balisage est donc tout à fait correct. Ce sont en effet des lexèmes dialectaux, pour lesquels les informations de datation et de sources bibliographiques sont fournies implicitement par les sources associées à l’étiquette géolinguistique dans le *Beiheft*.

Parmi les précisions reconnues ont été balisées, outre les sigles bibliographiques, dix datations explicites. Les références bibliographiques et les dates formant ensemble une fourchette de datation ont été balisées comme attendu au moyen de l'élement XML <attestation> (→ 5.4.20) :

<attestation><date>1640</date>—<biblio>Trév 1732</biblio></attestation>
<attestation><date>1640</date>—<date>1672</date></attestation>
<attestation><biblio>Rich 1680</biblio>–<biblio>Trév 1732</biblio>
</attestation>

Les informations nécessaires à une consultation transversale du FEW via les lexèmes (et via les molécules associées aux lexèmes) sont donc rendues accessibles.

Un type d'information manque toutefois dans la liste des cellules affichées en fin d'article pour leur assurer une autonomie complète en dimension T : il s'agit du marquage textuel. Certains des lexèmes ont en effet été classés comme des dérivés et certains comme des sens figurés. Cette information, qui relève de la dimension monographique mais qui peut intéresser une consultation transversale en dimension T, n'est pas descendue au niveau de la cellule dans le balisage tel qu'il se présente actuellement. Elle reste néanmoins accessible indirectement par le balisage de l'article en dimension M.

6.2.3 Informations non reconnues

Le balisage de l'article CHOCOLATL que nous venons d'examiner correspond à la modélisation proposée et répond déjà, en grande partie, aux besoins d'explicitation à la fois de la dimension monographique et de la dimension thesaurus du FEW. Il reste cependant à vérifier qu'aucune information n'a été, soit oubliée, soit balisée de façon incorrecte. Ces deux vérifications sont facilitées par les commentaires XML qui sont affichés en fin de document par les algorithmes de post-traitement.

6.2.3.1 Balisage manquant

Dans chacune des cellules, les molécules ont été reconnues et balisées de façon exhaustive. Le commentaire XML associé à l'algorithme *show-untagged-unit-text* indique en effet que la totalité des extraits de texte situés dans les cellules lexicales ont pu être associées à un type d'information :

< !– show-untagged-unit-text
[...]
Average tagging rate over units of the article : 100%
–>

Le texte non reconnu en dehors des cellules est affiché par l'algorithme *show-untagged-text*. Les trois paragraphes de documentation ont été balisés de façon exhaustive. Dans l'entrée, seule la glose de l'étymon n'est pas balisée, conformément à nos choix de modélisation (→ 3.5.3) :

< !– show-untagged-text
Untagged text in entry, column 63b (tagging rate : 51%)
<...> getränk aus kakao.

La couverture du balisage dans le commentaire et dans l'unique note de l'article

est très partielle (1/3 environ du total), ce qui est tout à fait normal puisque ces deux paragraphes contiennent un texte non structuré. L'affichage des parties non balisées permet de vérifier si un élément important a échappé aux algorithmes. La mention de langue *aztek.*, notamment, n'a pas été balisée dans le commentaire, en raison de son absence dans le *Beiheft* :

> Untagged text in <com> paragraph, column 63b (tagging rate : 36%)
> <...> ist aus <...> entlehnt, das selber aus<...>einem aztek. <...> übernommen ist. <...> ist<...>dieses selber erst im <...> ausdrücklich belegt, und<...>seine bildung innerhalb des aztek. ist nicht klar, <...>dazu Nykl <...>. Deutlich auf aztek.<...>ursprung weist der erste spanische beleg, von <...>der <...> geschrieben ist. Aus dem <...> auch <...> berichtet als<...> erster, dass das getränk auch in Spanien und Italien<...>genossen werde, und <...> schreibt :<...>Unklar ist auch das verhältnis von <...> zu <...> erste<...>beleg von <...> kommt von den kleinen Antillen, wo-<...>durch wahrscheinlich gemacht wird, dass diese form<...>sich selbständig verbreitet hat. Sie lebt auch in<...> Merkwürdi-<...>gerweise wird sie im <...> auch für das span. auf<...>den Philippinen bezeugt, unter der form <...>
> Untagged text in <notes> paragraph, column 64a (tagging rate : 31%)
> <...>1) Diese bed. ist eine synonymalablt. zu <...>, das vom volk <...> gesprochen wurde. <...>. noch<...>

6.2.3.2 Balisage erroné

La détection des erreurs de balisage est facilitée quant à elle par l'algorithme *show-tags*, qui affiche la liste des éléments balisés :

< !– show-tags

Tags inserted into entry, column 63b

chocolatl
<etymon>chocolatl</etymon>
<lang_etymon>(aztek.)</lang_etymon>

Tags inserted into <p> paragraph, column 63b

<pnum>1.</pnum>

<unit>Mfr. chocholate m. „breuvage fait avec des amandes de cacao" (1598)
</unit>
<geoling>Mfr.</geoling>
<form>chocholate</form>
<i>chocholate</i>
<gram>m.</gram>
<def>„breuvage fait avec des amandes de cacao"</def>
<precisions>(1598)</precisions>
<attestation>1598</attestation>
<date>1598</date>

<unit>chocolate (1640—Trév 1732)</unit>
<imp/>
<form>chocolate</form>
<i>chocolate</i>
<imp/>
<imp/>
<precisions>(1640—Trév 1732)</precisions>
<attestation>1640—Trév 1732</attestation>
<date>1640</date>
<biblio>Trév 1732</biblio>
[...]

Le parcours de cette liste permet de s'assurer rapidement que le balisage est correct. En ce qui concerne les paragraphes de documentation, aucune erreur n'apparaît. En revanche, dans le commentaire et les notes apparaissent visiblement quelques balises erronées : il s'agit d'une part de *s.* (« siehe ») devant un sigle bibliographique balisé comme catégorie grammaticale, d'autre part de l'article allemand *der* et de la conjonction *doch* balisés comme sigles bibliographiques. Ces erreurs peuvent être corrigées manuellement après rétroconversion, par simple suppression des balises erronées.

6.3 Rétroconversion du premier corpus de test

Les 130 articles composant le premier corpus de test sont les suivants :

– articles du projet DETCOL : ABLATIVUS (24, 34ab), ACCUSATIVUS (24, 94b), ACTIVUS (24, 115b-116b), ANTECEDENS (24, 642ab), APPELLATIVUS (25, 32b),

CASUALIS (2, 479b), COMPARARE (2, 970ab), COMPLETUS (2, 982b-983b), CONIUNCTIVUS (2, 1053ab), CŌPŬLA (2, 1157b-1162b), DEFECTIVUS (3, 28b), DEMONSTRARE (3, 38a), DEPONERE (3, 45a), DERIVARE (3, 49b), DISJŬNGERE (3, 96b-97a), FEMININUS (3, 451b-452a), FREQUENTATIO (3, 776b), FUTURUS (3, 929ab), GĔMĬNARE (4, 93ab), GENITIVUS (4, 102b), GĔNUS (4, 116a-117a), GĔRĔRE (4, 119a-120a), GRADUS (4, 204b-208b), GŬBĔRNARE (4, 299b-300b), HETEROS (4, 417b), ĬNCHOARE (4, 622b-623a), ĬNDĬCARE (4, 642a-644a), INFINITIVUS (4, 669a), INFINITUS (4, 669b), INTERJECTIO (4, 755b), ĬNTĔRRŎGARE (4, 761a-762a), MASCŬLĪNUS (6/1, 424b), NĔGARE (7, 82b-85a), NŌMEN (7, 175a-178b), OPTARE (7, 378b), PAENULTIMUS (7, 464b), PARTICIPIUM (7, 677a), PASSIVUS (7, 735b-736a), PĔRFĔCTUS (8, 237a-238a), PERSONALIS (8, 273a-274b), PLURALIS (9, 101ab), POSITIVUS (9, 231ab), PRAEPONERE (9, 302a), PRAESENS (9, 306b-308a), PRAETERIRE (9, 322b-323a), PRIMITIVUS (9, 380b-381a), PRONOMEN (9, 444b-445a), RECIPROCUS (10, 149b-150a), SUBSTANTIVUS (12, 357a), TEMPORALIS (13/1, 181ab), TRANSĪRE (13/2, 206a-209a), VOCATIVUS (14, 588b) ;

- articles supplémentaires du volume 3 : DRAWBACK, DREADNOUGHT, DRECK, DREG, DREK, DREO, DRESCHE, DREVE, DREVELEN, DRIBBLE, DRIESC, DRIFT-ICE, DRIL, DRILL, DRILL (-PLOUGH), DRILLEN, DRILL(ING), DRIN-, DRINGEN, DRINKEN, DRITS, DRIVE, DRIVE, DRIVER, DRIVING, DRŒSEM, DROGERIJ (3, 158a-160b) ;
- articles supplémentaires du volume 6 :
  - MABILLE, MACAIRE, * MACCARE, MACCHABAEUS, MACEDONIA, * MACELLARE, MACELLARIUS, MACELLUM, MACER, MACERARE, MACERIA (6/1, 1a-9b) ;
  - MAGOG, MAGOS, MAGUS, MAHON, MAHÓN, MAIĀLIS, MAIANUS (6/1, 51a-53b) ;
- articles supplémentaires du volume 16 : HARMSKARA, HARPA, HARPA, HÂRR, HARST, *HART, HART, HÄRTEN, HASAL, HASCHEN, HASE, HASELA, HASPA (16, 172a-176b) ;
- articles supplémentaires du volume 19 : BĀBŪŠ, BĀDĀM, BĀḎĀWARD, BADAWĪ, BĀḎ(A)RŪĞ (19, 15b-16b) ;
- articles supplémentaires du volume 21 : « colza », « lotier », « lupin », « mélilot », « pois de senteur, pois à bouquets », « sainfoin », « spergule », « tabac », « trigonelle » (21, 148a-149a) ;
- articles supplémentaires du volume 25 : AUGMENTATOR, AUGMENTUM, AUGSBOURG, AUGUR, AUGURACULUM, AUGURALIS (25, 882a-884b).

6.3.1 Temps de rétroconversion

L'implémentation a mené à des optimisations pour que les traitements coûteux (notamment les recherches de mots-clés appartenant à des listes très longues telles que la liste des étymons) prennent le moins de temps possible. En revanche, le temps de chargement des listes de mots-clés n'est pas compressible. Ce temps de chargement étant

constant quel que soit le nombre d'articles à rétroconvertir, il est plus avantageux de rétroconvertir un grand nombre d'articles en une fois.

Sur un portable Intel dual core[2] avec 4GB de RAM, la rétroconversion d'un seul article de taille moyenne (une colonne du FEW) s'effectue en environ deux à cinq secondes. La rétroconversion des 130 articles du premier corpus de test s'effectue en 138 secondes, ce qui revient à 1.06 seconde par article.

6.3.2 Analyse du balisage inséré

Les 130 articles du premier corpus comprennent ensemble 63 896 balises, parmi lesquelles 14 398 balises typographiques (FFML) et 37 672 balises sémantiques (FSML). Les 11 826 balises restantes sont des balises <provenance>, où se trouvent entre autres les 2 415 avertissements émis par les algorithmes. Les balises se répartissent comme suit :

```
FFML
   130 x <art>
   149 x <b>
   253 x <col>
   352 x <e>
   130 x <few>
     0 x <h>
  5326 x <i>
  6866 x <lb>
    65 x <notes>
   952 x <p>
  5403 x <provenance>
   175 x <sc>
 19801   balises FFML

FSML
    44 x <affix>
   283 x <appelnote>
  2843 x <attestation>
  4016 x <biblio>
   115 x <com>
    16 x <concept>
  1692 x <date>
  3392 x <def>
   524 x <doc>
   130 x <entry>
   528 x <etymon>
  4995 x <form>
  4676 x <geoling>
  1024 x <gram>
   223 x <group>
  3327 x <imp>
     0 x <k>
   601 x <lang>
    48 x <lang_etymon>
     0 x <milestone>
    31 x <mixt>
   281 x <note>
     3 x <part_com>
   328 x <pnum>
  2966 x <precisions>
   397 x <pref>
     0 x <provenance_exception>
  4008 x <provenance_update>
  2415 x <provenance_warning>
    31 x <renvoi>
   129 x <signature>
   231 x <title>
  4798 x <unit>
 44095   balises FSML
```

[2] Intel(R) Core(TM)2 Duo CPU T6500.

6.3.2.1 Informations très bien reconnues

Une grande majorité des balises sont correctement insérées, particulièrement en ce qui concerne les types d'informations suivants.

Etymons-vedettes. Les 121 étymons vedettes du corpus ont été reconnus, qu'ils se situent dans l'entrée ou qu'ils soient repris dans le commentaire de l'article dont ils constituent la vedette. L'étymon DRILL(ING) a toutefois été balisé partiellement, sans la partie entre parenthèses, ce qui constitue l'indice d'une non-correspondance exacte entre la liste *few-etymon-base* et les formes du FEW.

Langues d'étymon. Les langues d'étymon explicites (48 en tout) apparaissant dans les entrées des articles ont toutes été correctement reconnues et balisées. Tous les étymons-vedettes ont reçu un attribut *lang*, répétant la mention explicite ou rétablissant la langue implicite (il s'agit à chaque fois, dans notre corpus, du latin). Des 9 éponymes du corpus, 7 ont reçu la mention « nom-propre », excepté MAHON et MAHÓN, pour lesquels apparaissaient dans le FEW les mentions explicites (et plus précises) *PN* (pour *Personenname*) et *ON* (pour *Ortsname*).

Sous-lemmes. Si l'on excepte des cas de confusion entre les étymons et les lexèmes galloromans (voir ci-dessous), les sous-lemmes font partie des types d'information bien reconnus. 59 étymons ont été correctement reconnus comme sous-lemmes parmi les 130 articles, ce qui montre l'importance de leur balisage.

Concepts. Les 9 articles de concepts que compte le corpus ont été reconnus et correctement balisés comme tels. Les 7 concepts en surnombre sont les cibles de renvois internes, tous correctement balisés.

Signatures. Les articles de renvoi ne comportant pas de signature, les 129 signatures reconnues (dont 40 signatures explicites et 89 implicites) concernent 126 articles du corpus. Les trois signatures en surnombre proviennent de trois articles à signature multiple : AUGMENTATOR (*MJMathieu* et *Chauveau*), AUGMENTUM (*MJMathieu* et *Chauveau*) et CONIUNCTIVUS (*He* et *Lb*).

Documentation et commentaire. Le balisage des signatures a permis, comme attendu, de déterminer la métalangue de chaque article et de distinguer ainsi les paragraphes de commentaire, les paragraphes de documentation et les paragraphes mixtes. Des 952 paragraphes que compte le corpus, 945 (99, 3%) sont correctement balisés. Les 7 erreurs (0,7% du total) concernent six articles (AUGSBOURG, DEMONSTRARE, FREQUENTATIO, "lotier", "pois de senteur", "spergule"). Trois d'entre eux sont des articles de concept, dont on sait que le balisage est provisoire (→ 5.4.18).

Marquage de structuration. Le marquage alphanumérique est parfaitement reconnu dans toutes les parties de documentation balisées comme telles : des 330 marqueurs que contient le corpus, 328 ont été correctement balisés (<pnum>). Les deux marqueurs manquants apparaissent dans des paragraphes mixtes.

Notes et appels de note. Les appels de note sont reconnus de façon exhaustive. La différence entre les 281 notes (réparties sur 65 articles du corpus) et les 283 appels de note provient des articles MAIALIS et NOMEN, qui appellent une même note (respectivement la note 2 et la note 16) à deux reprises. Cette particularité,

tout à fait permise dans le discours fewien, est signalée comme attendue par l'attribut *call-sequence* de la balises <notes> de ces deux articles :

> <notes call-sequence="some-redundancy" count="18"> (FEW 7, 178a, NOMEN)
> <notes call-sequence="some-redundancy" count="3"> (FEW 6, 53a, MAIALIS)

Définitions. Les définitions (<def>) des parties documentaires du corpus sont toutes correctement balisées.

6.3.2.2 Informations reconnues avec des erreurs non problématiques

Les types d'information suivants sont bien reconnus également. Leur balisage présente quelques faux positifs ou quelques faux négatifs, le plus souvent non problématiques.

Référence au marquage de structuration. Les références au marquage de structuration dans les parties de commentaire (<pref>) sont parfaitement reconnues. Toutes les références contenues dans les commentaires des 130 articles ont été détectées : nous n'avons trouvé aucun faux négatif. Parmi les 397 références balisées se trouvent en revanche 22 faux positifs (6%), dont 16 concernent l'item *a*, particulièrement ambigu dans les commentaires en français (AUGMENTUM, MACCHABAEUS). Remarquons que le corpus contient 375 références correctes pour 328 marqueurs de documentation : la différence indique que le FEW peut faire plusieurs fois référence, dans un commentaire d'article, à un même marqueur de la documentation.

Marqueurs textuels. Les marqueurs textuels (<title>) sont très bien reconnus. Les 231 marqueurs balisés comportent seulement 5 faux positifs (2%), dont 3 sont des sigles géolinguistiques ou bibliographiques (non reconnus car présentant des variantes par rapport aux sigles répertoriés dans le *Beiheft*, voir ci-dessous). Huit éléments <title> contiennent un extrait de commentaire apparaissant dans un paragraphe mixte, tel que *Daraus entlehnt* (par exemple dans DEFECTIVUS). Le balisage des marqueurs textuels s'avère donc un moyen de reconnaissance des parties de commentaire dans les paragraphes mixtes. Les 218 balises <title> restantes contiennent des marqueurs textuels tout à fait classiques (→ 3.7.2).

Catégories grammaticales. Les 1 024 catégories grammaticales balisées contiennent 91 erreurs (9%), qui concernent uniquement les items *s* et *v*. Ces deux items sont en effet particulièrement ambigus, puisqu'ils introduisent fréquemment des expressions de renvoi bibliographique. Ces erreurs apparaissent uniquement dans des paragraphes de commentaire ou dans les parenthèses de précisions : elles n'ont, de ce fait, aucune incidence négative sur la détection des cellules lexicales dans les parties de documentation.

Langues. Le balisage des 602 mentions de langue ne présente aucun faux positif. En revanche, plusieurs mentions n'ont pas été reconnues, en raison de la non exhaustivité de la liste de mots-clés. Il s'agit de toponymes ou de sigles (tels que *altumbr.*, *apadov.*, *burgund.*, *obeng.*, *ostmndl.*, *spätlt.*, *westgerm.* etc.) non répertoriés dans le *Beiheft*.

Dates. Le balisage des 1 692 dates présente une dizaine de faux positifs, consistant en des références bibliographiques (numéros de page essentiellement). Ces erreurs sont explicables en partie par la non exhaustivité de la liste des références bibliographiques et en partie par une implémentation non encore parfaite de l'algorithme tag-renvoi (voir ci-dessous).

Attestations. La reconnaissance des attestations au sein des éléments de précision est très bonne : les 2 843 attestations comptent 16 erreurs (0,5%), dont 12 consistent en le balisage de cellules apparaissant au sein des précisions. Dans l'extrait suivant par exemple, les cellules contenant les lexèmes *coupleur*, *accoupler* et *accouplement* ont été balisées comme des attestations :

> Nfr. *couple* m. „système de 2 forces égales, parallèles, et de direction opposée, agissant aux 2 extrémités d'un levier" (seit 1863) ; „élément de pile électrique" (seit Lar 1907 ; dazu *coupleur* „appareil employé pour recharger les accumulateurs", seit Lar 1890 ; *accoupler* v. a. „grouper plusieurs générateurs électriques, pour augmenter leur rendement", seit Lar 1907 ; *accouplement* „réunion de plusieurs générateurs électriques", seit Lar 1890). (FEW 2, 1159b, CŌPŬLA I 2 a β)

Le balisage de ces cellules par l'élément <attestation> est bienvenu, car il permet de rassembler des éléments qui, sinon, seraient considérés comme indépendants. L'élément <attestation> joue finalement assez bien son rôle de regroupement, au sein des précisions, des informations qui fonctionnent ensemble. Cette opération de regroupement permet également de pallier des incohérences dans le balisage des sigles bibliographiques contenus dans les fourchettes de datation :

> <attestation> <biblio>Pom 1671</biblio>–<date> 1700</date>
> </attestation> ;
> <attestation> <biblio>Lar 1866–1948</biblio> </attestation> [FEW 24, 94b, ACCUSATIVUS]

Affixes. Sur les 44 affixes potentiels reconnus dans le corpus, 36 sont bien des affixes, et trois autres sont bien signalés comme ambigus (il s'agit à chaque fois d'un caractère isolé suivi d'un tiret). Parmi les cinq erreurs restantes, trois sont des marques dérivationnelles ou flexionnelles (*futurismo, -ista* ; *gesta, -orum* ; *négateur, -trice*) et deux sont des erreurs dues à des tirets en fin de ligne mal interprétés (*har-pagnement*, *mascu-lus*). La présence de ces cinq faux positifs ne pose pas de problème dans le processus de rétroconversion. Nous avons trouvé un seul faux négatif, non reconnu car il s'agit d'un préfixe appartenant également à la liste des étymons (*tri-*, FEW 7, 178a, NŌMEN).

Renvois. 38 renvois internes ont été balisés dans le corpus. Les 7 renvois conceptuels et les 4 articles de renvoi que contient le corpus ont été parfaitement reconnus et balisés comme tels. Les 18 renvois restants apparaissent dans le texte des articles étymologiques. Le balisage de ces derniers est encore perfectible, car il ne couvre pas la totalité de l'expression de renvoi.

6.3.2.3 Informations présentant des erreurs problématiques

Quelques types d'information présentent des erreurs problématiques. La plupart d'entre elles concernent des collisions entre deux types d'information. Certaines de ces ambiguïtés sont correctement signalées dans les balises <provenance> des algorithmes concernés.

Étymons et formes. Le corpus présente plusieurs cas de confusion entre des formes et des étymons en italiques : parmi les 75 étymons typés comme sous-lemmes, 14 sont des formes galloromanes et non des étymons (voir par exemple *hallo*, FEW 16, 175b, HASAL 1 c). Ces erreurs apparaissent essentiellement dans les parties de documentation, mais n'empêchent pas la reconnaissance des cellules lexicales : elles sont donc repérables par inspection des éléments <unit> dans le commentaire XML affiché en fin d'article par l'algorithme de post-traitement *show-tags*. La correction de ces erreurs est essentielle pour une utilisation du FEW en dimension thesaurus, car la non-reconnaissance d'une forme dans une cellule lexicale peut avoir des conséquences sur le rétablissement de l'implicite dans les cellules suivantes.

Précisions. Les 2 966 précisions balisées sont pour la plupart correctes. Dans les précisions sans parenthèses (→ 5.4.19) apparaissent quelques faux négatifs, dus à des références bibliographiques non détectées (en raison de la non exhaustivité de la liste des sigles bibliographiques). Parmi les précisions situées entre parenthèses, nous avons relevé 32 erreurs (1%), consistant toutes dans le balisage d'informations complémentaires placées entre parenthèses et insérées entre deux des quatre molécules obligatoires de la cellule lexicale (généralement entre le signifiant et le signifié). Il s'agit essentiellement d'informations grammaticales (*v.a.*) ou morphologiques (*f. trancha* ; *-abilis*). Ces erreurs, dont la correction est essentielle pour une délimitation correcte des cellules lexicales concernées, sont heureusement facilement repérables. Les 32 occurrences sont, en effet, bien signalées comme ambiguës dans la balise <provenance> affichée par tag-precisions.

Sigles géolinguistiques. Le balisage des 4 676 sigles géolinguistiques présente très peu de faux positifs (ceux que nous avons repérés consistent par exemple en un balisage erroné, dans le commentaire, de l'article allemand *die*, associé à tort au géonyme *Die*). Les quelques faux négatifs sont essentiellement dus à des collisions entre les listes de sigles bibliographiques et de sigles géolinguistiques (→ 5.4.9). Ces collisions sont, dans la plupart des cas, correctement résolues par leur balisage en tant que sigle bibliographique ; néanmoins, les quelques cas où ce balisage est erroné posent problème, car ils empêchent la détection d'une cellule lexicale. C'est le cas, dans le corpus, des sigles *Cr* (GŬBĔRNARE), *Orne* ("pois de senteur"), *Var* ("spergule") et *Gr Combe* (CASUALIS, GŬBĔRNARE, HÂRR, ĬNDĬCARE). La correction de ces ambiguïtés, signalées dans les balises <provenance> affichées en fin de document par l'algorithme *tag-biblio*, est donc essentielle.

Sigles bibliographiques. Les cas de collision mentionnés ci-dessus sont problématiques. Si l'on excepte ces cas, les 4 016 sigles bibliographiques balisés comme tels dans le corpus présentent quelques erreurs assez systématiques et peu problématiques. Ces erreurs concernent

- les deux mots allemands *der* et *doch* lorsqu'ils apparaissent en début de phrase ;
- la lettre s et la lettre v, lorsqu'elles introduisent un renvoi bibliographique et qu'elles sont séparées de ce renvoi par un mot tel que *noch* ou *auch* ;
- les lettres C, F, D, L, S, V, lorsqu'elles constituent un composant d'un mot ou d'une expression (par exemple dans *D'Urfé* ou *S'agit-il*) ;
- des mots apparaissant dans le texte du commentaire, tels que *France* et *Par* dans l'extrait suivant : *[...] assez répandues en France. Par abréviation, [...]* (AUGSBOURG).

Les faux négatifs existent, mais apparaissent comme peu nombreux relativement au nombre total de sigles balisés ; certaines références sont en outre balisées partiellement, comme *Brun* dans *Brun de Long Borc* (ANTECEDENS).

Cellules lexicales. Les erreurs rencontrées dans la délimitation des cellules lexicales sont dues soit à des faux positifs dans le balisage des précisions, soit à des confusions entre les formes et les étymons (voir ci-dessus). Ces erreurs mises à part, les cellules lexicales font partie des types d'information dont la reconnaissance est tout à fait satisfaisante. L'implicite y est correctement rétabli, excepté dans deux cas. D'une part, lorsque le signifiant est une locution, la catégorie grammaticale rétablie en fonction des cellules précédentes est évidemment erronée. D'autre part, les cellules non pourvues de précisions ne bénéficient pas du rétablissement des molécules situées après la dernière molécule explicite de la cellule. Le rétablissement de ces molécules doit encore être implémenté.

6.3.3 Conclusion

Nous n'écartons pas la possibilité de corriger de façon automatisée une partie des erreurs relevées dans le corpus. Plusieurs d'entre elles pourraient, en effet, être prévenues par l'écriture d'algorithmes plus complexes. Toutefois, leur correction amènera d'autres erreurs de balisage à d'autres endroits dans le texte fewien, qui nécessiteront à leur tour des modifications importantes dans d'autres algorithmes. Nous avons donc déterminé, en fonction des tests effectués sur le premier corpus et après quelques essais d'optimisation, une limite qui nous semblait raisonnable, dans l'optique, non de réduire à tout prix le nombre d'erreurs, mais de privilégier des options qui rendaient les erreurs aisément détectables. L'analyse des résultats de balisage montre que les erreurs vraiment problématiques sont assez typiques et délimitées pour être repérées et corrigées rapidement de façon manuelle. Par rapport à l'ensemble des informations correctement balisées, elles s'avèrent, en outre, assez peu nombreuses.

Malgré les erreurs de balisage, l'application des algorithmes sur le premier corpus assure la reconnaissance des éléments importants de l'entrée (étymons-vedettes, langues d'étymon, concepts et articles de renvoi). La structuration de l'article et les relations entre ses différents champs (appels de note, marquage alphanumérique, marquage textuel) est également assurée. La détection des cellules lexicales est assurée en grande partie également, les erreurs restantes étant détectables grâce aux commentaires XML affichés par *show-tags* et *show-untagged-unit-text*. La dimension monographique d'un article du FEW est donc explicitée comme attendu. La dimension thesaurus est quant à elle accessible dans les limites permises par le contenu des listes de mots-clés,

la non exhaustivité de ces listes étant la cause principale des faux négatifs détectés : ce problème est, par conséquent, soluble.

6.4 Rétroconversion du second corpus de test

Le second corpus de test est constitué de 16 articles, qui se divisent en trois parties :

- 10 articles concernés par le projet DETCOL : AORISTOS [24, 668b], IMPERATIVUS [4, 584b-585a], IMPERFECTUS [4, 586a], INDECLINABILIS [4, 637b], LAKONIKOS [5, 133a], NŌMĬNARE [7, 179a-182b], ORATIO [7, 385b-386b], *PASSARE [7, 707a-727b], PATER [8, 8a-13a], POSSESSIVUS [9, 237b-238a] ;
- 5 articles choisis de façon plus ou moins arbitraire dans le FEW : ATRIUM [25, 687b-691b], CALĒRE [2, 82b-85a], CHOCOLATL [2, 642b-643a], CHOCOLATL [20, 63b-64a], *SKĬRMJAN [17, 118a-120a] ;
- un article de la refonte hors volumes papier : *BASSIA, publié sur le site `www.atilf.fr/few`.

L'article de la refonte constitue un cas particulier, puisque le logiciel de rétroconversion n'a pas été écrit à son intention. Il a été ajouté au corpus de test à la demande des rédacteurs du FEW, intéressés par la possibilité de rétroconvertir les nouveaux articles[3]. Parmi les 15 articles restants se trouvent deux autres cas particuliers, volontairement choisis pour leur spécificité : d'une part, l'article ATRIUM, qui occupe une place particulière dans le balisage des définitions (→ 5.4.8), et d'autre part, l'article *PASSARE, qui nous intéresse parce que, outre son appartenance au corpus concerné par le projet DETCOL, il fait partie de l'ensemble des articles du FEW particulièrement longs (42 colonnes).

6.4.1 Temps de rétroconversion

L'inclusion, dans le second corpus de test, de l'article *PASSARE permet d'examiner le temps que prend la rétroconversion d'articles longs. Sur un portable Intel dual core[4] avec 4GB de RAM, la rétroconversion de l'article *PASSARE s'effectue en presque 5 minutes. Le temps de rétroconversion total du corpus (article de la refonte exclu) s'effectue en 295 secondes, ce qui revient à environ 20 secondes par article. La rétroconversion des 15 articles du second corpus nécessite donc davantage de temps que la rétroconversion des 130 articles du premier corpus. Ce résultat s'explique par le fait que la rétroconversion du seul article *PASSARE monopolise un cœur pendant 5 minutes et continue donc de s'exécuter bien après que les 14 autres articles ont été rétroconvertis. Si nous excluons *PASSARE du corpus, la rétroconversion des 14 articles restants s'effectue en 66.67 secondes, ce qui revient à 4.76 secondes par article et correspond aux résultats obtenus pour la rétroconversion du premier corpus.

[3]Les articles de la refonte de la tranche alphabétique B- sont, en effet, rédigés directement sur ordinateur, mais dans une optique de rédaction modulaire (cf. Matthey et Nissille 2010) qui ne permet pas directement le balisage de leur dimension monographique.

[4]Intel(R) Core(TM)2 Duo CPU T6500

6.4.2 Analyse du balisage inséré

Mises à part quelques erreurs inexplicables, clairement dues à des bogues dans l'implémentation des algorithmes (bogues qui seront donc à corriger avant que le logiciel ne soit mis en production sur l'ensemble du FEW), les 13 articles du second corpus que nous avons analysés (en mettant à part les articles ATRIUM, *BASSIA et *PASSARE) présentent les erreurs suivantes :

- des références bibliographiques et des étiquettes géolinguistiques non reconnues en raison de leur absence dans les listes de mots-clés (voir par exemple *Quimperfr.* dans FEW 7, 179b, NŌMĬNARE I). La plupart de ces erreurs ne sont pas problématiques. Un cas particulier est toutefois particulièrement remarquable : il s'agit du balisage comme sigle géolinguistique de *Thaon* dans la référence bibliographique (non reconnue, car s'étendant sur deux lignes) *Ph Thaon*, ce qui provoque des erreurs de rétablissement de l'implicite dans les cellules lexicales qui suivent (ces dernières étant étiquetées *Thaon* au lieu de *afr.*) :

  > — Afr. *numer* „énumérer“ Ph Thaon, *nomer* Tristan, *lommer* (hap. 13. jh.) (FEW 7, 179b, NŌMĬNARE I)

- le balisage comme sigles bibliographiques de lettres majuscules (*C*, *L*) ou de l'article allemand *der* (sous la forme *Der* en début de phrase) ;
- le balisage erroné de la lettre *s*, suivie d'un point, comme catégorie grammaticale dans une date ou devant un renvoi bibliographique ;
- le rétablissement d'une catégorie grammaticale implicite erronée après le signifiant d'une locution ;
- le balisage, comme précisions, d'informations complémentaires situées entre un signifiant et un signifié ;
- le balisage, comme attestations, de cellules lexicales enchâssées dans les précisions d'une autre cellule ;
- le balisage erroné, comme attestation, de numéros de volume et de page introduits par un point-virgule après un sigle bibliographique ;
- le regroupement de deux (ou trois) cellules en une seule à la suite d'un *id.* non balisé.

Ces erreurs apparaissent de façon ponctuelle et sont peu nombreuses par rapport à l'ensemble des balises insérées (26 339 balises sur la totalité des 15 articles, dont 8 450 balises FFML, 24 179 balises FSML et 2 775 balises <provenance>).

6.4.3 Conclusion

L'analyse du second corpus de test confirme les résultats du premier corpus. Les erreurs de balisage constatées correspondent en effet aux erreurs déjà détectées dans le premier corpus. Aucun nouveau type d'erreur n'a été relevé, ce qui est rassurant. Le second corpus étant néanmoins assez limité, et le FEW étant un dictionnaire rempli de

surprises, il serait illusoire d'affirmer qu'aucune erreur de type nouveau n'apparaîtra lors de la rétroconversion de la totalité du FEW. On peut toutefois supposer que les différents commentaires XML affichés en fin d'article permettront aisément la détection d'éventuelles erreurs non constatées dans les corpus de test.

6.5 Modalités d'exploitation des articles rétroconvertis

Le balisage XML inséré par le logiciel ne suffit pas, à lui seul, pour permettre une exploitation informatisée du FEW. La rétroconversion des articles constitue la deuxième des trois étapes menant à l'informatisation du FEW (→ 4.1) : l'étape suivante consiste à mettre au point un moteur de recherche capable de lire les documents rétroconvertis et de répondre aux requêtes des utilisateurs, ainsi qu'une interface adéquate de consultation et de lecture du FEW informatisé. Par ailleurs, l'informatisation du *Beiheft* est nécessaire pour permettre l'explicitation des nombreux sigles et abréviations du FEW.

La création d'un *Beiheft* informatisé, d'un moteur de recherche et d'une interface de consultation et de lecture sortent du cadre de cette thèse ; néanmoins, la plupart des fonctionnalités abordées dans les chapitres précédents dépendent directement de ces outils. Nous nous permettons donc ci-dessous quelques réflexions et propositions qu'il nous semble important de prendre en compte pour permettre une exploitation optimale des articles rétroconvertis. Nous aborderons successivement l'informatisation du *Beiheft*, les modalités de consultation du FEW informatisé en dimension monographique et les modalités de lecture du FEW en dimension thesaurus. Dans la foulée, nous nous permettrons ensuite quelques réflexions concernant la mise à jour du FEW informatisé.

6.5.1 Informatisation du *Beiheft*

Durant le processus de rétroconversion, les listes d'abréviations et de sigles contenues dans le *Beiheft* servent d'indicateurs de reconnaissance pour les types d'information suivants : sigles géolinguistiques, sigles bibliographiques et mentions de langues non galloromanes. Le balisage de ces types d'information est essentiellement destiné à fournir trois fonctionnalités :

1. la possibilité d'une explicitation complète de ces sigles et abréviations lors de la lecture d'un article (dimension monographique) ;
2. des consultations transversales du FEW selon ces types d'informations (rechercher par exemple tous les lexèmes du francoprovençal) ;
3. des consultations transversales du FEW selon des types d'information implicites (rechercher par exemple tous les lexèmes attestés à partir du 18e siècle, quelle que soit la façon – date explicite ou référence bibliographique – dont le FEW renseigne cette datation).

Ces trois fonctionnalités nécessitent obligatoirement l'informatisation du *Beiheft*, sous forme de base de données exploitable par un moteur de recherche. Une telle base de données doit idéalement contenir :

– une table des sigles géolinguistiques, accompagnés

    - de leur explicitation,
    - d'une localisation à différents niveaux structurés (si pertinent),
    - des références bibliographiques (*sources*) qui leur sont directement associées ;
- une table des sigles bibliographiques, accompagnés
    - de leur explicitation (référence complète),
    - de renseignements complémentaires éventuels,
    - d'une localisation à différents niveaux structurés (si pertinent),
    - d'une datation à différents niveaux structurés (si pertinent) ;
- une relation directe entre les sigles bibliographiques des deux tables.

Le contenu de la base de données correspondrait ainsi aux informations telles qu'elles sont données dans le *Beiheft*, excepté pour les champs *localisation* et *datation* de la table bibliographique, qui sont des ajouts destinés à faciliter le travail du moteur de recherche. Par exemple, le sigle géolinguistique *Fourn.* serait accompagné des champs suivants :

- explicitation : Fournels
- localisation : FR, Lozère, Mende, Fournels
- sources : 1. ALLo — 2. CampsAtlas p 13

Le sigle bibliographique *ALLo* serait quant à lui accompagné des champs suivants :

- explicitation : Hallig, Rudolf, *Atlas linguistique de la Lozère et des cantons limitrophes du Gard et de l'Ardèche*
- renseignements complémentaires : (copie des 2 485 cartes manuscrites en possession du FEW ; les enquêtes furent faites, de 1932 à 1934, dans 35 points par R. Hallig en Lozère, R. Böhne dans le Gard et H. Brendel en Ardèche ; les données sans référence à un numéro de carte proviennent de fiches supplémentaires)
- localisation : FR, Lozère — FR, Gard — FR, Ardèche
- datation : 20e siècle, 1932–1934

Une relation directe serait établie entre le sigle *ALLo* apparaissant dans le champ des sources de la table des sigles géolinguistiques et le sigle *ALLo* de la table des sigles bibliographiques.

L'utilisation d'une telle base de données serait la suivante.

Lors de la lecture d'un article en dimension monographique, le contenu du champ *explicitation* (cf. ci-dessus) est fourni à l'utilisateur par un simple clic sur un sigle géolinguistique ou bibliographique. La possibilité lui est fournie de développer l'explicitation. S'il s'agit d'un sigle bibliographique, les renseignements complémentaires

sont affichés. S'il s'agit d'un sigle géolinguistique, la localisation et les sources canoniques sont affichées, ces dernières sous forme de lien hypertextuel renvoyant à leur explicitation dans la table des sigles bibliographiques.

Lors de la consultation du FEW en dimension thesaurus, les contenus structurés des champs *localisation* et *datation* sont utilisés par le moteur de recherche pour répondre à des requêtes portant sur des critères géographiques ou chronologiques. Par exemple, l'information *Lozère* présente dans le champ *localisation* du sigle *Fourn.* permet au moteur de recherche de retrouver un lexème rattaché dans le FEW à l'étiquette géolinguistique *Fourn.* lors d'une requête concernant la Lozère. De même, les informations présentes dans le champ *datation* du sigle *ALLo* (cf. ci-dessus) permettent de considérer comme appartenant à la période contemporaine les cellules lexicales présentant le sigle *ALLo*, ainsi que celles présentant l'étiquette géolinguistique *Fourn.* (ou toute autre étiquette géolinguistique qui contient *ALLo* dans le champ de ses sources canoniques). Le contenu de ces champs *localisation* et *datation*, ainsi que la mise en relation des deux tables bibliographique et géolinguistique, doivent donc permettre au moteur de recherche de passer outre l'implicite du discours fewien.

L'utilisation en dimension thesaurus du *Beiheft* informatisé est, évidemment, dangereuse si elle n'est pas encadrée. Deux précautions sont à prendre, l'une par le moteur de recherche et l'autre par l'utilisateur.

– Le moteur de recherche doit respecter certaines règles. L'une d'elles est la priorité de l'étiquette géolinguistique sur le sigle bibliographique. Il n'est pas pertinent, par exemple, d'utiliser le sigle *ALLo* pour une recherche sur la Lozère, puisque ce sigle vaut aussi pour une partie du Gard et de l'Ardèche. Une requête géographique doit donc s'effectuer uniquement sur les informations présentes dans la table des sigles géolinguistiques.

– L'utilisateur doit quant à lui tenir compte des divers niveaux auxquels se situent les résultats retournés par le moteur de recherche. En effet, les lexèmes retournés en réponse à une requête sur la Lozère peuvent ne pas couvrir toute la Lozère, mais certaines localités seulement du département. Il est donc de la responsabilité de l'utilisateur (aidé éventuellement par une interface pensée dans cette optique) de vérifier les étiquettes géolinguistiques attribuées à chacun des lexèmes reçus.

Dès lors, nous pensons qu'il serait pertinent, lorsque l'utilisateur émet une requête sur un critère géographique, de lui proposer un choix entre deux possibilités : soit une recherche portant sur les mentions explicites du FEW (qui retournerait comme résultats uniquement les lexèmes dont l'étiquette géolinguistique est *Lozère*), soit une recherche prenant en compte l'implicite fewien (qui utiliserait les données contenues dans les champs *localisation* du *Beiheft* informatisé). Dans le second cas, un avertissement serait émis en tête de la liste des résultats, signalant le caractère hétérogène de ceux-ci et l'importance de contextualiser les données par une lecture attentive du FEW. Une telle remarque s'applique également, *mutatis mutandis*, aux consultations transversales concernant un critère chronologique.

6.5.2 Modalités de consultation

Si nous mettons à part les cas particuliers de consultation transversale évoqués ci-dessus qui mettent en cause l'implicite fewien et en l'occurrence le *Beiheft*, les modalités de consultation des articles rétroconvertis dépendent directement des balises XML qui sont insérées dans ces derniers. Nous n'énumérerons pas en détail chacun des types de requêtes susceptibles d'être mis à disposition des utilisateurs : les besoins de ces derniers ont déjà été évoqués plus haut (→ 2.4.4), et le balisage reflète assez fidèlement ces besoins, puisqu'il a été pensé notamment en fonction d'eux. Nous voudrions néanmoins résumer l'essentiel de ce qu'il nous semble souhaitable de fournir à l'utilisateur en termes d'interface (types d'interrogations, présentation des résultats) et émettre quelques considérations pratiques sur les éléments qu'il sera nécessaire de prendre en compte lors de la mise au point du moteur de recherche.

6.5.2.1 Types d'interrogations

Conformément aux besoins exprimés par les utilisateurs, l'interface de recherche doit proposer au moins :

- une recherche plein texte (sans prise en compte du balisage inséré) ;
- une recherche (simple) de lexèmes (à entrer tels quels par l'utilisateur, avec une fonction de recherche qui neutralise les diacritiques (→ 4.2.2), ou à choisir dans une liste des formes du FEW) ;
- une recherche (simple) d'étymons (à entrer tels quels par l'utilisateur, avec une fonction de recherche qui neutralise les diacritiques (→ 4.2.2), ou à choisir dans une liste des étymons du FEW) ;
- une recherche (simple ou complexe) de types d'information précis, tels qu'un sigle géolinguistique, un sigle bibliographique, un marqueur textuel, une langue d'origine etc. (à choisir dans les listes correspondant à chaque type d'information).

6.5.2.2 Moteur de recherche

Le moteur de recherche doit évidemment déterminer, pour chaque type de requête proposé à l'utilisateur, dans quelles balises il doit chercher l'information fournie et quel extrait de texte doit être retourné comme résultat.

Prenons l'exemple d'un utilisateur voulant effectuer une recherche sur les lexèmes du francoprovençal. Ce type de recherche implique une requête portant sur les sigles géolinguistiques (cette option « recherche d'un sigle géolinguistique » doit donc être proposée à l'utilisateur, sur la base d'une liste des sigles permis). Les balises concernées sont les balises <geoling> : le moteur de recherche passe en revue le contenu de toutes les balises <geoling>[5] et retient celles qui contiennent la mention *frpr.* (mot choisi par l'utilisateur dans la liste déroulante des sigles géolinguistiques). Le résultat

[5]En pratique, la recherche du mot *frpr.* dans les balises <geoling> ne se fera certainement pas directement dans le texte du FEW, mais dans une table d'indexation préalablement créée de façon automatique (cf. Dendien 1996, à propos du moteur de recherche du TLFi).

fourni est (par exemple) la liste des cellules lexicales contenant les éléments <geoling> retenus, avec lien hypertextuel vers l'article du FEW où se trouvent ces cellules.

L'élaboration d'un moteur de recherche capable de lire le texte balisé du FEW pour répondre aux interrogations des utilisateurs sort du cadre de cette thèse. Néanmoins, nous nous permettons d'insister sur l'importance, lors de requêtes complexes, d'une définition correcte de la *portée* de chaque type d'information.

Considérons les exemples de requête complexe suivants :

1. le sigle géolinguistique *wall.* ET le sigle bibliographique *Rem*[2] ;
2. le sigle géolinguistique *wall.* ET le marqueur textuel *ablt.* ;
3. le sigle géolinguistique *wall.* ET une langue d'étymon germanique.

Les types d'interrogation mis en jeu dans ces trois requêtes n'ont pas la même portée. La portée d'un sigle géolinguistique ou d'un sigle bibliographique (comme de toute autre molécule) est la cellule lexicale (<unit>) dans laquelle il se trouve. La portée d'un marqueur textuel est le paragraphe (<p>) lorsqu'il débute celui-ci ou, s'il apparaît au milieu d'un paragraphe, le groupement de lexèmes (<group>) dont il fait partie. Enfin, la portée d'une langue d'étymon est l'article complet (<art>).

Pour répondre correctement aux trois requêtes, le moteur de recherche doit prendre en compte la portée du type d'information le plus large. Pour la première requête, les deux informations doivent appartenir au même élément <unit>. Les deux informations de la deuxième requête doivent appartenir au même élément <group> ou au même élément <p> (selon que le marqueur est inclus ou non dans un élément <group>). Enfin, les deux informations de la troisième requête doivent apparaître dans le même élément <art>.

6.5.2.3 Présentation des résultats

Les résultats fournis par le moteur de recherche doivent, nous semble-t-il, correspondre exactement à un extrait de texte du FEW, accompagné d'un lien hypertexte vers l'article complet. Il n'est donc pas question de réécrire les articles ou de présenter les données autrement que le fait le FEW, comme l'ont proposé certains (→ 1.3).

– La recherche simple d'un étymon-vedette (<etymon type="vedette">), d'une langue d'étymon (<etymon lang="[...]">) ou d'un concept (<concept>) doit retourner la totalité de l'élément <entry> (entrée de l'article) dans lequel ont été trouvées ces informations, avec lien hypertextuel vers l'article complet.

– La recherche simple ou complexe d'informations moléculaires (<biblio>, <date>, <def>, <form>, <geoling>, <gram>, <imp/>) doit retourner la totalité de l'élément <unit> (cellule lexicale) dans lequel ont été trouvées ces informations, avec lien hypertextuel vers l'article complet.

– La recherche simple d'informations susceptibles de se trouver n'importe où dans le FEW, notamment hors entrée et cellules lexicales (<affix>, <biblio>, <date>, <title>, <signature>), doit retourner un extrait plus ou moins large du texte fewien dans lequel se trouve l'information (la largeur, à déterminer de façon exacte

en fonction de chaque type d'information, étant définie comme X caractères textuels à gauche et Y caractères textuels à droite de l'élément XML contenant l'information), avec lien hypertextuel vers l'article complet.

6.5.3 Modalités de lecture

Une fois sa recherche effectuée, l'utilisateur est renvoyé vers des articles complets du FEW. Le balisage inséré par la rétroconversion doit faciliter la *lecture* de ces articles ; pour ce faire, il est important que l'interface de lecture du FEW informatisé offre à l'utilisateur les deux fonctionnalités que sont, d'une part, la mise en évidence d'informations particulières (parmi lesquelles la structuration de l'article) et, d'autre part, des liens hypertextuels à divers niveaux.

6.5.3.1 Mise en évidence d'informations particulières

Le plan sommaire de l'article fourni en tant que commentaire XML (→ 6.2.1 ; 5.4.24) peut être affiché par l'interface de lecture, de façon à fournir à l'utilisateur une vision synthétique du contenu de l'article. De même, le marquage alphanumérique devrait être mis en évidence dans l'article de façon très visible.

L'interface devrait en outre proposer de mettre en évidence dans l'article (par exemple au moyen d'un surlignage coloré) certains types d'information (par exemple, tous les étymons cités dans l'article) ou certains mots (par exemple, toutes les mentions *frpr.*), au choix de l'utilisateur. Si l'article est affiché à la suite d'une requête portant sur certains types d'information particuliers, ces derniers seront mis en évidence de la même façon dans le texte de l'article.

6.5.3.2 Liens hypertextuels

Les liens hypertextuels suivants doivent être mis en place[6] :

– dans un même article, entre le marquage alphanumérique de la documentation (<pnum>) et les références à ce marquage dans le commentaire (<pref>) ;

– dans un même article, entre les appels de note (<appelnote>) et les notes (<note>) ;

– entre un renvoi (<renvoi>) et un autre article (via l'étymon-vedette (<etymon> dans <entry>) ou le numéro de page (<col>), selon le contenu du renvoi) ;

– entre un sigle (<geoling>, <biblio>) et le *Beiheft* informatisé ;

– entre un sigle (<biblio>) et une ressource externe.

[6]Des observations sur la mise en place d'un système cohérent de liens hypertextuels ont été formulées dans Briquet et Renders à paraître.

6.5.4 Modalités de mise à jour du FEW

La rétroconversion des 25 volumes du FEW est censée permettre leur mise à jour. Cette thèse a permis de relever divers avis quant à cette question délicate (→ 2.4.3). Voici ce qu'on peut en conclure.

1. L'informatisation du FEW *doit* permettre sa mise à jour, comprise comme l'intégration des corrections et ajouts déjà publiés dans la littérature annexe, depuis des années, par des experts reconnus.
2. L'informatisation du FEW *peut* permettre que ce dernier devienne évolutif, c'est-à-dire qu'il continue à intégrer de nouvelles propositions de corrections et ajouts non publiées.
3. L'intégration des corrections et ajouts, quel que soit le moyen à l'aide duquel elle s'opèrera en pratique, devra respecter certaines règles :
   (a) les corrections et ajouts devront être dûment signalés comme tels, afin qu'il soit toujours possible de faire la distinction entre le FEW original et une version corrigée du FEW ;
   (b) les corrections et ajouts devront être datés ;
   (c) les corrections et ajouts devront être vérifiés par un (ensemble d')expert(s) avant publication et doivent être signés par leur auteur, qui en est responsable.
4. Un FEW informatisé et consultable sur le web doit donner accès à la fois au FEW original et au FEW mis à jour. Ce dernier sera référencé par l'adresse du site internet et la date de consultation. En outre, un système de *versioning* est envisageable, qui permettrait, d'une part, de conserver de façon rigoureuse la trace de toutes les modifications effectuées dans le FEW et, d'autre part, de définir, malgré une mise à jour permanente, des version stables du dictionnaire consultables par le grand public.

La rétroconversion des articles provoquera en outre inévitablement des corrections typographiques mineures, telles que l'ajout d'une parenthèse ou d'un guillemet manquant. Les erreurs du FEW détectées durant la saisie des articles ou durant le processus de rétroconversion (par exemple durant l'application des algorithmes de prétraitement) doivent, en effet, être corrigées pour permettre la rétroconversion de l'article. Durant les tests effectués dans le cadre de cette thèse sur les deux corpus mentionnés ci-dessus, la trace du texte original a été provisoirement conservée sous forme de commentaires XML, insérés à l'endroit exact de la correction et expliquant la modification effectuée, de la façon suivante :

> <– [corr] commentaire –>

Dans l'extrait suivant par exemple (FEW 4, 205a, GRADUS I 1 a), une parenthèse fermante a été ajoutée :

> afrb. <i>greaux</i> (15. jh., Z 24, 213)< !– [corr] fixed FEW : added closing parenthesis –>.

Ce procédé n'est pas optimal, puisqu'il ne respecte pas tous les principes cités ci-dessus. Dès qu'un outil de *versioning* sera rendu disponible pour la mise à jour du FEW, il sera nécessaire d'y intégrer également ces corrections.

6.6 Conclusion

La mise au point des outils qui permettront l'exploitation finale des articles rétroconvertis sort du cadre de cette thèse. En ce qui concerne la rétroconversion elle-même, nous pensons avoir abouti à un résultat globalement très satisfaisant. Le balisage automatiquement inséré dans les deux corpus de test par le logiciel de rétroconversion comporte certes encore des erreurs, mais les commentaires XML et les avertissements destinés à signaler les cas problématiques pallient en grande partie ces imperfections en permettant leur repérage.

Dans l'optique, qui a été la nôtre, de préférer un grand nombre d'erreurs visibles à un petit nombre d'erreurs cachées, il nous semble que l'utilisation des nombreuses listes de mots-clés s'est révélée particulièrement intéressante. La non exhaustivité de ces listes (inhérente à la grande variabilité du discours fewien) est, visiblement, la cause d'une grande partie des erreurs rencontrées. L'utilisation de ces listes apparaît donc à première vue très critiquable. À la question de savoir si d'autres moyens de détection, non dépendants de listes, n'auraient pas été préférables, nous pouvons répondre par la négative, et cela en nous appuyant sur deux arguments. Tout d'abord, les listes dont nous disposons s'avèrent, en fin de compte, ne pas conduire à un nombre d'erreurs beaucoup plus important que d'autres indicateurs. Ensuite, l'utilisation de listes rend les résultats du logiciel facilement améliorables par simple ajout de mots-clés dans les fichiers concernés, ce qui ne nécessite aucune intervention dans le code informatique du logiciel. Les mots-clés manquants étant pour la plupart aisément repérables grâce à l'algorithme de post-traitement *show-untagged-text*, il est possible à un linguiste rétroconvertissant un article de les relever et de les ajouter aux listes de mots-clés. Cette possibilité d'améliorer le résultat de balisage sans devoir connaître le code informatique sous-jacent au logiciel est très confortable. Par ailleurs, une fois qu'un nouveau mot-clé a été inséré, il est disponible pour tout article à rétroconvertir par la suite. Nous pouvons dès lors espérer que la nécessité de compléter les listes diminuera au fur et à mesure de l'avancement de la rétroconversion du FEW.

Malgré les erreurs (et d'autant plus si ces dernières sont corrigées à mesure qu'elles sont reconnues), les articles rétroconvertis sont en grande partie conformes à la modélisation proposée. À partir du moment où les outils d'exploitation appropriés seront mis au point, le balisage XML inséré de façon automatique dans les articles du FEW devrait permettre l'utilisation du dictionnaire dans ses deux dimensions et répondre ainsi à la majorité des besoins exprimés par les utilisateurs.

Chapitre 7

Conclusion

Au terme de cette étude, il apparaît que la problématique de départ était mal posée. Il est possible d'informatiser le FEW, sans nul doute. La question de la faisabilité du projet cache une autre question, plus importante : quelle forme doit prendre cette informatisation, sachant qu'elle doit à la fois répondre aux attentes qu'elle suscite et rester réalisable dans les limites imposées par les particularités du discours fewien ? Sachant qu'elle doit respecter le discours fewien tel qu'il a été construit, sans chercher à le modifier outre mesure, et malgré tout ouvrir la voie à de nouvelles possibilités d'exploitation ?

Nous espérons avoir répondu à cette question. La modélisation du discours fewien que nous proposons à la fin de la première partie de cette étude a été définie de façon à respecter trois objectifs, parfois contradictoires : faciliter la consultation et la lecture de l'ouvrage, refléter ses structures et être applicable de façon automatisée. Le premier objectif a nécessité une compréhension précise des problèmes rencontrés par les utilisateurs du FEW et de leurs attentes. La perspective, qui a été la nôtre, d'appréhender les structures de l'ouvrage en les observant du point de vue de l'utilisateur a permis de mettre en évidence une dimension double du FEW, perçu comme un thesaurus d'unités lexicales d'une part, comme un recueil de monographies d'autre part. Ces deux façons d'appréhender l'objet FEW sont toutes deux justifiables, mais nuisibles l'une à l'autre si elles ne sont pas considérées dans une relation d'interdépendance étroite.

La prise en compte de cette double dimension dans la modélisation du FEW passe par la résolution, dans certaines limites, de l'implicite fewien. En rétablissant l'autonomie, d'une part de l'article (dimension monographique), d'autre part de l'unité lexicale (dimension thesaurus), nous ouvrons la voie à de nouveaux parcours de consultation du dictionnaire tout en ne dénaturant pas la cohérence du discours fewien tel qu'il a été construit durant la longue période de sa rédaction.

Le modèle proposé étant formalisé en XML, son application dans le texte fewien consiste à insérer des balises autour de chaque type d'information pertinent. La densité de l'ouvrage oblige à ce que ce balisage soit effectué de façon automatisée. La deuxième partie de notre étude, en examinant les moyens d'automatiser l'insertion du balisage XML dans le discours fewien, a permis de tester les limites d'un processus automatique de rétroconversion. Ces limites sont moins sévères que nous ne le pensions de prime abord. À condition que le texte de chaque article du FEW soumis au proces-

sus ait d'abord été saisi en respectant certaines normes de codage et de mise en forme, la plupart des types d'information définis dans le modèle s'avèrent reconnaissables de façon automatisée. Les combinaisons d'indicateurs utilisées par les algorithmes que nous avons conçu à cet effet sont assez spécifiques et fiables pour réduire le nombre d'erreurs. La reconnaissance d'un type d'information nécessitant souvent le balisage préalable d'un autre type d'information, l'ordre d'application des algorithmes joue un grand rôle dans le succès de l'opération. Une partie des erreurs restantes pourra être évitée par une étude plus approfondie de certains algorithmes ; les autres erreurs, qui peuvent être corrigées manuellement, sont repérables grâce aux avertissements émis par les algorithmes de balisage. Des commentaires XML affichés dans les documents rétroconvertis aident également à la détection de ces erreurs.

Les nombreuses incohérences du discours fewien, qui faisaient craindre l'impossibilité d'une informatisation, ne constituent donc pas un obstacle infranchissable. En définitive, le discours fewien reste assez consistant dans sa variation pour être décrit selon des règles utilisables par une machine, avec une marge d'erreur assez faible. La présence d'erreurs marque toutefois une limite dans ce qu'il est raisonnable d'espérer d'un traitement totalement automatisé.

La rétroconversion des 25 volumes du FEW est donc possible sous la forme de documents XML, correspondant chacun à un article, dans lequel tous les types d'information considérés comme pertinents sont reconnus de façon automatisée, avec une marge d'erreur corrigible manuellement. Le résultat de cette opération constitue ce qu'il est raisonnable d'attendre du processus dans les limites permises par le discours fewien. Nous considérons toutefois ce résultat, non comme une fin, mais comme une première étape. Il pourra, par la suite et si la communauté scientifique le juge bénéfique, être enrichi manuellement de façon à mieux encore expliciter le discours lexicographique particulier du FEW et permettre son exploitation raisonnée.

Le projet d'informatisation du FEW peut être mis en route. Le logiciel conçu dans le cadre de cette étude est opérationnel et utilisable pour rétroconvertir la totalité des articles des 25 volumes. Lorsqu'un plus grand nombre d'articles que celui que compte notre corpus de test aura été soumis au logiciel et rétroconverti, il sera possible de vérifier la robustesse du processus et d'effectuer les améliorations algorithmiques qui ne manqueront pas de se révéler nécessaires. Les tâches à réaliser pour permettre l'accomplissement de l'opération de rétroconversion sont :

– l'acquisition des articles du FEW selon les normes de codage et de mise en forme (balisage typographique) définies pour le format d'entrée du logiciel ;
– l'amélioration des algorithmes après rétroconversion d'un grand nombre d'articles, de façon à réduire encore le nombre d'erreurs ;
– le stockage des articles rétroconvertis sur un serveur et la mise au point d'outils (moteur et interface de recherche) permettant de les exploiter sans attendre que la totalité du FEW ait été traitée.

Avant de clore définitivement notre étude, qu'il nous soit permis d'insister sur deux points qui ont fortement contribué à l'aboutissement de notre démarche et de notre modeste contribution à cet ambitieux projet.

Réfléchir, avant de se lancer dans une entreprise d'informatisation, à l'objet lexicographique et à l'utilisation qu'on veut en faire nous semble un premier aspect détermi-

nant. Il est important de définir précisément le résultat attendu, en prenant en compte toutes les contraintes susceptibles d'influencer ce résultat. « [I]nformatisez un mauvais dictionnaire, le résultat sera un mauvais dictionnaire informatisé », disait Robert Martin (Martin 2001, 12), insistant ainsi sur le fait que l'informatisation n'était pas un remède miracle. La question qui nous a préoccupée dans cette thèse fut plutôt : « Informatisons un bon dictionnaire, quel sera le résultat ? » L'opération consistant à modéliser un objet lexicographique en vue de son informatisation n'est pas neutre (cf. Renders 2009). Une modélisation reste une interprétation, parmi d'autres possibles et tout aussi valables, de l'objet. Il est donc erroné de penser qu'il n'y a qu'une seule informatisation possible ; le résultat du processus est toujours perfectible et sujet à critiques.

Le second point sur lequel nous voudrions insister est l'intérêt, dans un projet de cette ampleur, d'une collaboration interdisciplinaire étroite. Cette étude n'est pas le fruit de deux travaux isolés, dont l'un serait l'écriture des algorithmes (réalisée par un linguiste) et l'autre leur implémentation (réalisée par un informaticien). Le mécanisme des chaînes virtuelles ($\rightarrow$ 4.4.2.3), qui a suscité l'intérêt des spécialistes du langage XML (cf. Briquet et Renders 2010), n'a pu être conçu, implémenté et intégré dans les algorithmes qu'à partir du moment où une collaboration étroite s'est mise en place entre les deux disciplines. De même, l'analyse sur corpus des résultats de balisage et la recherche de solutions aux erreurs rencontrées n'ont pu s'effectuer qu'en joignant une connaissance approfondie du FEW et des algorithmes à une connaissance précise du code Java utilisé dans le logiciel. La construction du logiciel présenté dans cette thèse a été possible parce que nous avons bénéficié de la collaboration d'un excellent informaticien. Cette collaboration a un inconvénient : la communication entre les deux disciplines n'est pas facile, et un certain temps doit lui être consacré. Mais le résultat en vaut la peine.

Tant qu'une réflexion linguistique approfondie ira de pair avec des solutions informatiques efficaces, nous pensons que l'informatisation d'un dictionnaire ne peut être que bénéfique pour la communauté de ses utilisateurs. Nous espérons que notre modeste contribution aura eu le mérite de poser les bases réflexives d'un projet de rétroconversion qui nécessitera encore, en raison des particularités et des incohérences du FEW, de régulières mises au point

Liste des sigles bibliographiques

ATILF = Analyse et Traitement Informatique de la Langue Française

Beiheft = Wartburg, Walther von, 1950[2] [1929[1]]. *Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes. Beiheft : Ortsnamenregister, Literaturverzeichnis,Übersichtskarte*, Tübingen : Mohr.

Beiheft Supplement = Hoffert, Margarete, 1989 [1957]. *Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes. Supplement zur 2. Auflage des Bibliographischen Beiheftes*, Bâle : Zbinden.

BlochWartburg = Bloch, Oscar / Wartburg, Walther von, 1968[5] [1932[1]]. *Dictionnaire étymologique de la langue française*, Paris : PUF.

Complément = Chauveau, Jean-Paul / Greub, Yann / Seidl, Christian, 2010[3]. *Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes. Complément*, Strasbourg : Éditions de linguistique et de philologie (Bibliothèque de Linguistique Romane, Hors Série 1).

DAG = Baldinger, Kurt (dir.), 1975–. *Dictionnaire onomasiologique de l'ancien gascon*, Tübingen : Niemeyer.

DAO = Baldinger, Kurt (dir.), 1974–. *Dictionnaire onomasiologique de l'ancien occitan*, Tübingen : Niemeyer.

DEAF = Baldinger, Kurt (dir.), 1974–. *Dictionnaire étymologique de l'ancien français*, Tübingen : Niemeyer.

DHFQ = Poirier, Claude (dir.), 1998. *Dictionnaire historique du français québécois*, Québec : P.U. Laval.

DMF1 = ATILF / Équipe "Moyen français et français préclassique", 2003–2005. *Dictionnaire du Moyen Français* (DMF1). Nancy : Nancy-Université/CNRS, site internet (http ://www.atilf.fr/blmf).

DOM = Stempel, Wolf-Dieter (dir.), 1987–. *Dictionnaire de l'occitan médiéval*, Tübingen : Niemeyer.

DRF = Rézeau, Pierre (dir.), 2001. *Dictionnaire des régionalismes de France*, Bruxelles : De Boeck.

DSR = Thibault, André, 2004[2][1997[1]]. *Dictionnaire suisse romand*, Genève-Carouge, Zoé.

FEW = Wartburg, Walther von *et al.*, 1922-2002. *Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes* (25 vol.), Bonn/Heidelberg/Leipzig-Berlin/Bâle : Klopp/Winter/Teubner/Zbinden.

LEI = Pfister, Max[/Schweickard (Wolfgang)] (dir), 1979–. *Lessico Etimologico Italiano*, Wiesbaden : Reichert.

REW3 = Meyer-Lübke, Wilhelm, 1935^{3}[1911^{1}]. *Romanisches Etymologisches Wörterbuch*, Heidelberg : Carl Winter.

TLF = Imbs, Paul (dir.), 1971–1994. *Trésor de la langue française. Dictionnaire de la langue du XIXe et du XXe siècle (1789-1960)* (16 vol.), Paris : Éditions du CNRS/Gallimard.

TLFi = CNRS/Université Nancy2/ATILF, 2004. *Trésor de la Langue Française informatisé* (cédérom), Paris : CNRS Éditions (version internet : http ://stella.atilf.fr/).

Bibliographie

Amann, Bernd / Rigaux, Philippe, 2002. *Comprendre XSLT. Transformation de documents XML*, Cambridge / Cologne / Paris : O'Reilly.

Aprile, Marcello, 2004. *Le Strutture del* Lessico Etimologico Italiano, Galatina : Mario Congedo (Publicazioni del Dipartimento di filologia linguistica e letteratura dell'Università di Lecce, 24).

Arveiller, Raymond, 1999. *Addenda au FEW XIX (Orientalia)*, Tübingen : Niemeyer.

ATILF, 2003. *Französisches Etymologisches Wörterbuch. Index A–Z*, Paris : Champion.

Atkins, B.T. Sue / Rundell, Michael, 2008. *The Oxford Guide to Practical Lexicography*, New York : Oxford University Press.

Baldinger, Kurt, 1968. « Fr. *laie, layer*, die germanische Waldwirtschaft und eine neue Etymologie », *in* Stimm, Helmut / Wilhelm, Julius (éd.), *Verba et Vocabula, Ernst Gamillscheg zum 80. Geburtstag*, München : Wilhem Fink, 49–56.

——, 1974. « Le FEW de Walther von Wartburg. Introduction », *in* Baldinger, Kurt (éd.), *Introduction aux dictionnaires les plus importants pour l'histoire du français*, Strasbourg (= Paris, Klincksieck), *Bulletin des Jeunes Romanistes*, vol. 18-19, 11–47.

——, 1980. « Etymologies doubles dans le FEW », *in Italic and Romance : Linguistic Studies in Honor of Ernst Pulgram*, Amsterdam : Benjamin, 189–194.

——, 1988. *Etymologien, Untersuchungen zu FEW 21-23*, vol. I, Tübingen : Niemeyer.

——, 1993. « Vers une typologie des fautes dans le FEW : le redoublement des étymologies, des articles et des attestations », *in* Pickens, Rupert T. (éd.), *Studies in Honor of Hans-Erich Keller. Medieval French and Occitan Literature and Romance Linguistics*, Kalamazoo / Michigan : Western Michigan University, 507–532.

——, 1998. *Etymologien, Untersuchungen zu FEW 21-23*, vol. II, Tübingen : Niemeyer.

——, 2003. *Etymologien, Untersuchungen zu FEW 21-23*, vol. III, Tübingen : Niemeyer.

Beckert, Virginie, 2003. *L'informatisation du* Französisches Etymologisches Wörterbuch, Mémoire de DESS (industries de la langue, mention traitement automatique et techniques de traduction), Université de Metz, Metz.

Boutier, Marie-Guy, 2008. « Cinq relations de base pour traiter la matière géolinguistique : Réflexions à partir de l'expérience de l'Atlas linguistique de la Wallonie. », *Estudis Romànics* 30 : 301–310.

Boutier, Marie-Guy / Brochard, Marie-José / Büchi, Eva, 1990. « Cas d'étymologie double dans le FEW (III) », *Travaux de linguistique et de philologie* 28 : 25–36.

Boutier, Marie-Guy / Brochard, Marie-José / Büchi, Eva / Chambon, Jean-Pierre / Chauveau, Jean-Paul / Dondaine, Colette, 1992. « Cas d'étymologie double dans le FEW (IV) », *Travaux de linguistique et de philologie* 30 : 387–415.

Boutier, Marie-Guy / Büchi, Eva / Chambon, Jean-Pierre / Chauveau, Jean-Paul, 1994. « Cas d'étymologie double dans le FEW (V) », *Travaux de linguistique et de philologie* 32 : 37–68.

Briquet, Cyril / Renders, Pascale, 2010. « A Virtualization-Based Retrieval and Update API for XML-Encoded Corpora », in *Proceedings of Balisage : The Markup Conference 2010.* Balisage Series on Markup Technologies, vol. 5 (2010).

——, à paraître. « Une approche reposante (RESTful) des aspects opérationnels de la rétroconversion du Französisches Etymologisches Wörterbuch (FEW) », in *Actes de la première journée liègeoise de Traitement des Sourcesgalloromanes (TraSoGal) (Liège, 28 mai 2009).*

Brochard, Marie-José / Chambon, Jean-Pierre / Chauveau, Jean-Paul / Hoffert, Margarethe, 1989. « Cas d'étymologies doubles dans le FEW (II) », *Travaux de linguistique et de philologie* 27 : 151–179.

Büchi, Eva, 1991. « Contribution à l'étude des déonomastiques galloromans : Index des éponymes dans le FEW », *Nouvelle Revue d'Onomastique* 17/18 : 139–152.

——, 1992. « Le traitement des déonomastiques dans le FEW », *in* Hilty, Gerold (éd.), *Actes du XXe Congrès International de Linguistique et Philologie Romanes*, Tübingen : Francke, vol. IV, 69–78.

——, 1996. *Les Structures du 'Französisches Etymologisches Wörterbuch'. Recherches métalexicographiques et métalexicologiques*, Tübingen : Niemeyer.

Büchi, Eva / Chambon, Jean-Pierre, 1995. « Un des plus beaux monuments des sciences du langage : le FEW de Walther von Wartburg (1910-1940) », *in* Antoine, Gérard / Martin, Robert (éd.), *Histoire de la langue française, 1914-1945*, Paris : CNRS Editions, 935–963.

Chambon, Jean-Pierre, 1989a. « Aspects philologiques et linguistiques dans la refonte du FEW : utilité d'une approche métaphilologique des représentations linguistiques », *in* Kremer, Dieter (éd.), *Actes du 18e Congrès International de Linguistique et Philologie Romanes*, Tübingen : Niemeyer, vol. 7, 218–230.

——, 1989b. « Tradition et innovations dans la refonte du FEW », *in* Kremer, Dieter (éd.), *Actes du 18e Congrès International de Linguistique et Philologie Romanes*, Tübingen : Niemeyer, vol. 7, 327–337.

——, 1997. « Les emprunts du français moderne aux dialectes ou patois : une illusion d’optique en lexicologie française historique ? », *in Lalies. Actes des sessions de linguistique et de littérature*, Paris : Presses de l’école normale supérieure, vol. 17, 33–53.

Chambon, Jean-Pierre / Chauveau, Jean-Paul / Dondaine, Colette / Rézeau, Pierre, 1999. « Cas d’étymologie double dans le FEW (VI) », *Travaux de linguistique et de philologie* 37 : 251–262.

Chambon, Jean-Pierre / Degeller, Marie-José / Hoffert, Margaretha, 1987. « Cent cas d’étymologie double dans le FEW », *in* Tamba, Irène (éd.), *Études de lexicologie, lexicographie et stylistique offertes en hommage à Georges Matoré*, Université Paris-Sorbonne, 165–184.

Chauveau, Jean-Paul, 2006. « D’un site informatique en chantier pour le FEW », *in Nuovi media e lessicografia storica. Atti del colloquio in occasione del settantesimo compleanno di Max Pfister*, Tübingen : Niemeyer, 33–37.

Dendien, Jacques, 1996. « Access to information in a textual database : access function and optimal indexes », *in Research in Humanities Computing, Papers from the ACH-ALLC Conference*, Oxford : Clarendon Press.

Dendien, Jacques / Pierrel, Jean-Marie, 2003. « Le Trésor de la langue française informatisé. Un exemple d’informatisation d’un dictionnaire de langue de référence », *Traitement automatique des langues (TAL)* 43 (2) : 11–37.

Dijkstra, E. W., 1976. *A Discipline of Programming*, Prentice-Hall.

Gebhardt, Karl, 1982. « L’apport des dialectes d’oïl (surtout entre 1300 et 1600) au lexique de la langue commune (d’après le FEW) », *in* Wunderli, Peter (éd.), *Du mot au texte. Actes du IIIe Colloque International sur le Moyen Français*, Tübingen : Narr, 31–48.

Goosse, André, 1991. « La date de décès des mots », *in Mélanges de grammaire et de lexicologie françaises*, Louvai-la-neuve : Peeters, 163–177.

Hallig, Rudolf / von Wartburg, Walther, 1952. *Begriffssystem als Grundlage für die Lexikographie : Versuch eines Ordnungsschemas*, Berlin : Akademie-Verlag.

Harold, E.R. / Means, W.S., 2000. *XML in a nutshell : a desktop quick reference*, ACM.

Hartmann, R.R.K, 2001. *Teaching and Researching Lexicography*, Harlow : Pearson Education.

Kilgariff, Adam, 2005. « Informatique et dictionnairique », *Revue française de linguistique appliquée* : 95–102.

Kleene, S.C., 1956. « Representation of events in nerve nets », *in Automata Studies*, Princeton University Press, 3–40.

Lagueunière, France, 1998. « Le traitement de la variation diatopique en français moderne dans le Französisches Etymologisches Wörterbuch », *in Atti del XXI Congresso Internazionale di Linguistica e Filologia Romanza*, Tübingen : Niemeyer, vol. III, 387–395.

Malkiel, Yakov, 1976. *Etymological Dictionaries. A Tentative Typology*, Chicago-London : The University of Chicago Press.

Martin, Robert, 1996. « Sur l'implicite dans le langage ordinaire : la notion de non-dit », *Le Français Moderne* 64 : 129–135.

——, 2001. *Sémantique et automate. L'apport du dictionnaire informatisé*, Paris : Presses Universitaires de France.

——, 2004. *Comprendre la linguistique. Épistémologie élémentaire d'une discipline*, Paris : Quadrige / PUF, deuxième éd..

Matthey, Anne-Christelle, 2004. « La refonte des articles des étymons en B- du FEW : trouvailles et interrogations », Rencontres FEW – DEAF, ATILF, 5-6 novembre 2004, texte non publié.

Matthey, Anne-Christelle / Nissille, Christel, 2010. « L'irruption de l'informatique dans la rédaction du FEW », *in Actes du XXVe Congrès International de Linguistique et de Philologie Romanes (Innsbruck, 3-8 septembre 2007)*, Tübingen : Niemeyer.

Matthey, Anne-Christelle / Thibault, André, en préparation. « Le FEW de Walther von Wartburg : introduction pratique à son utilisation », Manuel.

Mazziotta, Nicolas, 2011. « L'informatisation du Französisches Etymologisches Wörterbuch. Concepts pour une approche modélisée commune à l'Atlas Linguistique de la Wallonie », *Zeitschrift für Romanische Philologie* 127 : 36–62.

Mazziotta, Nicolas / Renders, Pascale, 2010. « Vers un enrichissement raisonné de la rétroconversion du Französisches Etymologisches Wörterbuch (FEW) », *in* Dykstra, Anne / Schoonheim, Tanneke (éd.), *Proceedings of the XIV Euralex International Congress (Leeuwarden, 6-10 July 2010)*, Ljouwert : Fryske Akademy.

Merk, Georges, 1980. « Mots fantômes ou obscurs. Datations douteuses », *Revue de Linguistique Romane* 44 : 266–303.

Pfister, Max, 1980. *Einführung in die Romanische Etymologie*, Darmstadt : Wissenschaftliche Buchgesellschaft.

Plouzeau, May, 2000. « Compte rendu de Jean-Paul Chauveau, Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanisches sprachschatzes, von Walther von Wartburg. Fascicule no 157, tome XXV (refonte du tome 1er) : **auraticus—auscultare*, pp. 961-1056, Bâle, 1988 », *Revue de linguistique romane* 64 : 508–540.

Polguère, Alain, 2008. *Lexicologie et sémantique lexicale. Notions fondamentales*, Montréal : Les Presses de l'Université de Montréal, deuxième éd..

Quemada, Bernard, 1991. « Acquis et perspectives de l'informatique », *Travaux de linguistique* 23 : 17–21.

Renders, Pascale, 2009. « Des dangers de l'informatisation d'un document : le cas du FEW », *in Pratiques du document*, *Méthodes et Interdisciplinarité en Sciences humaines*, vol. 2, 179–195, `http://popups.ulg.ac.be/MethIS/document.php?id=263`.

Rey, Alain, 1971. « Le dictionnaire étymologique de W. von Wartburg : structures d'une description diachronique du lexique », *Langue française* 10 : 83–106.

Rey-Debove, Josette, 1971. *Etude linguistique et sémiotique des dictionnaires français contemporains*, Paris : Mouton.

Roques, Gilles, 1991. « L'articulation entre étymologie et histoire de la langue », *in Où en sont les études sur le lexique ? Bilan et perspectives*, De Boeck Université, *Travaux de linguistique*, vol. 23, 91–95.

Rouleux, Sophie, 2005. « La DTD et son langage XML. Une application pour la lexicographie contemporaine », *éla* 137 : 73–94.

Rézeau, Pierre, 1989. « Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanisches sprachschatzes, von Walther von Wartburg. Fascicule no 149, tome XXV (refonte du tome 1er) : **artificialis—aspergere*, pp. 385–480, Bâle, 1988 », *Cahiers de lexicologie* 54 : 165–168, compte rendu de Chambon (Jean-Pierre).

Schmitt, Hans Joachim, 1992. « Recherches lexicales sur des documents français issus du Refuge vaudois en Allemagne », *in* Hilty, Gerold (éd.), *Actes du XXe Congrès International de Linguistique et Philologie Romanes*, Tübingen : Francke, vol. IV, 307–320.

Stubblebine, Tony, 2003. *Regular Expression Pocket Reference*, Sebastopol (CA) : O'Reilly.

Swiggers, Pierre, 1990. « Lumières épistolaires sur l'histoire du F.E.W. : Lettres de Walther von Wartburg à Hugo Schuchardt », *Revue de linguistique romane* 54 : 347–358.

Tannier, Xavier, 2006. « Traiter les documents XML avec les *contextes de lecture* », *Traitement Automatique des Langues* 47 (1) : 61–85.

Van der Vlist, Eric, 2002. *XML Schéma*, Paris : O'Reilly.

von Wartburg, Walther, 1959. « Remarques sur les mots français dans le dictionnaire de M. Corominas », *Revue de linguistique romane* 23 : 207–260.

——, 1961. « L'expérience du FEW », *in Lexicologie et lexicographie françaises et romanes. Orientations et exigences actuelles (Strasbourg 12-19 novembre 1957)*, Paris : Éditions du CNRS, 209–219.

Wooldridge, Terence Russon, 1990. « Le FEW et les deux millions de mots d'Estienne-Nicot : deux visages du lexique français », *Travaux de linguistique et de philologie* 28 : 239–316.

——, 1998. « Le lexique français du XVIe siècle dans le GDFL et le FEW », *Zeitschrift für romanische Philologie* 114 : 210–257.

Zamboni, Alberto, 1976. *L'Etimologia*, Bologna : Zanichelli.

Annexe A

Balisage de la refonte

Cette annexe contient la DTD élaborée par Anne-Christelle Matthey et Gilles Souvay pour la refonte de la tranche alphabétique B- du volume 1 du FEW, telle qu'elle est présentée par ses auteurs. Auparavant, nous fournissons au lecteur non initié à XML quelques explications sur ce langage et sur la façon dont se présente une DTD.

A.1 Le langage XML[1]

XML est un langage de balisage structurel, qui permet de faire apparaître la structure sous-jacente d'un document informatique et les diverses informations qu'il offre au lecteur. Chaque balise, représentée entre les symboles < et >, peut être comparée à une étiquette qui, collée sur un récipient, indique son type de contenu. Dans le cas d'un dictionnaire, les types de contenus sont par exemple la vedette de l'article, la catégorie grammaticale et la définition, auxquels peuvent s'ajouter un exemple, un commentaire etc. Ces informations sont autant d'éléments qui seront entourés de leur balise respective. À titre d'exemple, un poème constitué d'un titre et de trois strophes pourrait être balisé ainsi :

<poeme>[2]
<titre> . . . </titre>
<strophe> . . . </strophe>
<strophe> . . . </strophe>
<strophe> . . . </strophe>
</poeme>

Comme on le voit, les balises forment des parenthèses autour des fragments de texte ainsi isolés. On distingue les balises ouvrantes <poème>, <titre>, <strophe> et les balises fermantes </poeme>, </titre>, </strophe>. Une paire de balises ouvrante et fermante et le fragment qu'elles entourent constituent ce qu'on appelle un élément XML. Le nom de la balise (par exemple poeme) est le type de l'élément[3]. Il est important de ne pas confondre le type de contenu, spécifié par le nom des balises (types d'éléments), et le contenu lui-même, qui se trouve entre les balises : tous les éléments

[1]Nous renvoyons également à Amann et Rigaux 2002 (chapitres 2 et 3), à Rouleux 2005 et à l'introduction à XML dans le guide en ligne de la TEI (*Text Encoding Consortium*).

[3]Pour cette terminologie XML et ce qui suit, voir Amann et Rigaux 2002, 15.

balisés <poeme> dans un recueil de textes ne comporteront évidemment pas le même poème.

Afin de garder une cohérence, il convient à la fois de définir pour chaque document les différentes balises qu'on souhaite y appliquer et de spécifier leur enchaînement logique. Il peut s'avérer nécessaire de spécifier que dans un poème, l'élément « titre » arrive toujours en premier lieu, que l'élément « strophe » peut apparaître plusieurs fois de suite et qu'il est lui-même composé d'un nombre fini d'éléments « vers ». De la même manière, dans un dictionnaire, l'élément « exemple » aura une place déterminée par rapport à l'élément « définition » et pourra (mais ne devra pas obligatoirement) comprendre un élément « références », lui-même constitué par exemple d'un élément « auteur » suivi d'un élément « titre ».

La définition des balises et de leur hiérarchisation est consignée dans un fichier annexe, appelé « DTD » (*Document Type Definition*) ; il s'agit d'une description générale de la structure du document. La DTD nomme les différents éléments et spécifie le contenu de chacun d'eux, de la façon suivante :

```
< !ELEMENT poeme (titre ?,strophe+)>
< !ELEMENT titre (#PCDATA)>
< !ELEMENT strophe (#PCDATA)>
```

La première de ces trois lignes déclare qu'il existe un élément de type « poeme » et, dans la parenthèse, le définit comme contenant deux sous-éléments, « titre » et « strophe ». Le contenu des éléments de type « titre » et « strophe » est spécifié dans les deux lignes suivantes comme étant une chaîne de caractères (#PCDATA, signifiant *Parser Character Data*).

En plus de déclarer, définir et hiérarchiser les éléments, la DTD permet de spécifier leurs conditions d'apparition. Elle utilise pour cela quelques signes de ponctuation, qui possèdent dans la DTD une signification syntaxique particulière.

1. En ce qui concerne l'ordre des éléments :

   – l'emploi de la virgule (,) entre des éléments signifie que l'ordre de ces éléments est fixé et doit correspondre à la séquence indiquée ;

   – l'emploi de la barre verticale (|), au contraire, indique que l'apparition de l'un ou de l'autre des éléments en question est obligatoire [ou seulement possible ?].

2. En ce qui concerne le nombre d'occurrences d'un élément ou d'un groupe d'éléments :

   – l'emploi du symbole + à droite d'un élément (ou d'un groupe d'éléments) signifie que ce dernier peut apparaître de une à n fois : on dira qu'il peut être répété, tout en gardant à l'esprit que ce qui est répété n'est pas le contenu (le poème en lui-même), mais le type de contenu (un autre poème, dans une anthologie par exemple) ;

   – dans la même position, le symbole * signifie que l'élément (ou le groupe d'éléments) peut apparaître de zéro à n fois ;

   – dans la même position, le symbole ? signifie que l'élément (ou le groupe d'éléments) peut apparaître de zéro à une fois : il sera dit optionnel.

L'expression citée plus haut signifie donc que dans l'élément « poeme », l'élément « titre » est optionnel, qu'il doit obligatoirement apparaître avant l'élément « strophe » et que ce dernier peut apparaître de une à n fois.

A.2 La DTD de la refonte

```
<!-- ============================================================ -->
<!--                                                              -->
<!-- FEW  : DTD pour un article du FEW                            -->
<!--                                                              -->
<!-- Anne-Christelle MATTHEY                                      -->
<!-- Gilles SOUVAY                                                -->
<!--                                                              -->
<!-- ATILF-CNRS/UniversitéNancy2/UHP                Fichier 2005 -->
<!--                                                              -->
<!-- ============================================================ -->

<!ELEMENT FEW  (Article)+>

<!ELEMENT Article  (FormesHereditaires, FormesSavantes?, FormesEmpruntees?, Commentaire?)>

<!ELEMENT FormesHereditaires  (NiveauHierarchique+)>
<!ELEMENT FormesSavantes      (NiveauHierarchique+)>
<!ELEMENT FormesEmpruntees    (#PCDATA)>
<!ELEMENT Commentaire         (#PCDATA)>

<!ELEMENT NiveauHierarchique  (NumeroTitre+, Paragraphe+)>
<!ELEMENT NumeroTitre         (Numero, Titre)>
<!ELEMENT Numero              (#PCDATA)>
<!ELEMENT Titre               (#PCDATA)>

<!ELEMENT Paragraphe          (Introduction?, (LexicographieFrancaise | Dialectes)+)>

<!ELEMENT LexicographieFrancaise (Langue+)>
<!ELEMENT Dialectes              (UnDialecte+, AutresReferencesD?)>

<!ELEMENT Introduction        (#PCDATA)>

<!ELEMENT Langue              (Derive?, (Etat, Etat?), ListeFormesL)>
<!ELEMENT ListeFormesL        (UneFormeL+)>
<!ELEMENT UneFormeL           (Forme, (CodeGrammatical?, Definition?, References),
                              (((CodeGrammatical, Definition?)|Definition),
                              References)*)>

<!ELEMENT Etat                (#PCDATA)>
<!ELEMENT Forme               (#PCDATA)>
<!ELEMENT CodeGrammatical     (#PCDATA)>
<!ELEMENT Definition          (#PCDATA)>

<!ENTITY % reference1         "(Date, Date?, Ref)" >
<!ENTITY % reference2         "(Fourchette, Source?)" >
```

```
<!ENTITY % reference3         "(Ref)" >

<!ELEMENT References          ((%reference1;) | (%reference2;) | (%reference3;))+>
<!ELEMENT Fourchette          ((Date, Ref) | (Ref, (Date | Ref)))>
<!ELEMENT Ref                 (CommentaireSource?, Source+, Secondaire?)>
<!ELEMENT CommentaireSource   (#PCDATA)>
<!ELEMENT Source              (#PCDATA | EXP)+>
<!ELEMENT Secondaire          (#PCDATA)>
<!ATTLIST Secondaire          Verifiee  (oui | non) "non">

<!ELEMENT Date                (#PCDATA)>

<!-- ============================================================= -->
<!--                                                               -->
<!-- ============================================================= -->

<!ELEMENT UnDialecte          (Derive?, Localisation+, ListeFormesD)>
<!ELEMENT ListeFormesD        (UneFormeD+)>
<!ELEMENT UneFormeD           (Forme, (CodeGrammatical | Definition | References)*)>
<!ELEMENT Localisation        (#PCDATA)>


<!ELEMENT Derive              (CommentaireDerive?, confixe)>
<!ELEMENT CommentaireDerive   (#PCDATA)>
<!ELEMENT confixe             (#PCDATA)>
<!ATTLIST confixe             Type  (suffixe | prefixe | infixe) #REQUIRED>

<!ELEMENT AutresReferencesD   (Introduction?, Ref+)>

<!-- ============================================================= -->
<!--                                                               -->
<!-- ============================================================= -->

<!ELEMENT EXP              (#PCDATA)>
```

Annexe B

Questionnaire

Cette annexe contient le questionnaire sur le FEW, tel qu'il a été diffusé en 2007. Le lecteur trouvera dans le chapitre 2 de cette étude une présentation et une analyse des réponses reçues.

Questionnaire FEW

Ce questionnaire comporte quatre parties. Les deux premières s'appliquent au FEW tel qu'il est et cherchent à répondre aux questions suivantes :

1. Comment l'informatisation peut-elle aider à la lecture d'un article complet du FEW ?
2. Comment l'informatisation peut-elle aider à la recherche d'une information ou d'un ensemble d'informations dans la totalité du FEW ?

Toujours dans le but de faciliter la consultation de ce dictionnaire, la troisième partie ouvre la porte aux « rêves fous » susceptibles de permettre une autre utilisation des matériaux enregistrés et analysés par le FEW.

Enfin, dans la perspective d'un dialogue entre les utilisateurs et les rédacteurs du FEW, la dernière partie cherche à savoir comment l'informatisation peut aider à la mise à jour de l'ouvrage.

Répondez aux questions posées en cochant les cases et en ajoutant des exemples et commentaires éventuels dans les zones pointillées prévues à cet effet.

Nom : ....

Prénom : ....

Adresse électronique : ....

Adresse postale : ....

Institution de rattachement : ....

I. Aide à la lecture du FEW

Vous paraît-il utile d'accéder directement (par « clic ») à des informations telles que :

☐ La résolution d'une abréviation géo-historique, bibliographique, de structuration etc. (ex. liég. → liégeois, ablt. → ableitungen « dérivés »)

....

☐ La ou les source(s) détaillée(s) correspondant à une abréviation géo-historique ou bibliographique

....

☐ La localisation d'un sigle géolinguistique (appartenance à un groupe linguistique supérieur : ex. liég. → wallon → oïl)

....

☐ Autres :

....

Vous paraît-il utile de mettre certaines informations mieux en évidence, par exemple

☐ Le plan de l'article

....

☐ Les étymons cachés

....

☐ Autres :

....

II. Aide à la recherche dans le FEW

Que cherchez-vous fréquemment dans le FEW ?

☐ L'étymologie d'un lexème particulier (« adresse FEW »)

- ☐ ancien / moyen français
- ☐ français moderne
- ☐ dialectal (langue d'oïl)
- ☐ occitan / gascon
- ☐ francoprovençal

....

☐ La date d'un lexème particulier

....

☐ Le sens d'un lexème particulier

....

☐ Un ensemble de mots (galloromans) ayant une caractéristique particulière

- ☐ Une localisation : ex. tous les mots liégeois enregistrés par le FEW

....

- ☐ Une source : ex. tous les mots de Trév 1771 enregistrés par le FEW

....

- ☐ Une signification : ex. tous les mots définis par « faim »

....

- ☐ Autres :

....

☐ Un ensemble de mots (galloromans) ayant une étymologie particulière

☐ Une langue-origine : ex. tous les mots d'origine bretonne

....

☐ Une caractéristique morphosyntaxique, sémiotique : ex. tous les noms venant d'un nom de lieu

....

☐ Autres :

....

☐ Un ensemble de mots (galloromans) ayant plusieurs caractéristiques particulières (« ET »)
Ex. tous les mots picards datés du 13e siècle

Exemple 1 :...

Exemple 2 :...

Exemple 3 :...

☐ Un ensemble de mots (galloromans) ayant l'une des caractéristiques particulières contenues dans un ensemble (« OU »)
Ex. tous les mots wallons (étiquetés « liég. » ou « nam. » ou …)
Ex. tous les mots moyen français (étiquetés « mfr. » ou datés dans la période considérée)

Exemple 1 :...

Exemple 2 :...

Exemple 3 :...

☐ Autres :

....

III. Rêves fous[1]

Quels projets, permettant une utilisation nouvelle du FEW, vous paraissent utiles ?

☐ La traduction en français des commentaires allemands

....

☐ Une cartographie automatique, permettant de visualiser sur la carte du domaine galloroman les données localisées d'un article (ou d'une partie d'article sélectionnée)

....

☐ Des liens avec d'autres documents informatisés (ou en voie de l'être) que vous consultez parallèlement au FEW lors de vos recherches

☐ Dictionnaires (TLFi, DEAF, DMF, …) :

....

☐ Bases de données :

....

☐ Autres :

....

☐ Autres :

....

IV. Aide à l'amélioration du FEW

Si vous pouviez effectuer des apports au FEW, quels seraient-ils ?

☐ Ajout d'une information (datation, source, localisation, … ?)

....

☐ Ajout d'un lexème

....

☐ Autres :

....

Si vous pouviez effectuer des corrections dans le FEW, quelles seraient-elles ?

☐ Correction de la datation d'un lexème particulier (antédatation, rétrodatation, postdatation)

....

☐ Correction de l'étymologie d'un lexème ou d'un ensemble de lexèmes

....

☐ Réécriture d'une partie d'article

....

☐ Autres :

....

Où voudriez-vous que se fasse la publication des corrections proposées ?

☐ Directement dans le FEW

....

☐ En dehors du FEW, mais regroupées en un seul endroit

....

Par qui ?

☐ Par celui qui propose la correction

....

☐ Par soumission à un expert (rédaction du FEW) qui validerait et intégrerait la correction

....

Quand ?

☐ Ponctuellement, au fur et à mesure des propositions

....

☐ À intervalles réguliers (lesquels ?)

....

Commentaires éventuels ?

Envoyer le formulaire

[1] Ces divers projets sont ouverts à candidature : avis aux traducteurs, cartographes et informaticiens intéressés !

Annexe C

Table des caractères du FEW

C.1 Introduction

Le document présenté ci-dessous fait l'objet d'une explication au chapitre 4.2.2.

C.2 Table des caractères

FEW Retroconversion character map (version 20101104) - Authored by Pascale Renders

Glyph	FFML & keyword encoding		FSML encoding	Flattening	Delimiter	Greek	Phonetic
	XML escape sequence	UTF-8	Unicode	ASCII	Flag	Flag	Flag
ă *ă*	ă		U+0103	a	.	n	y
ā *ā*	ā		U+0101	a	.	n	y
á *á*	á		U+00E1	a	.	n	y
Á	Á		U+00C1	A	.	n	y
à		à	U+00E0	a	.	n	y
â		â	U+00E2	a	.	n	y
ä *ä*		ä *ä*	U+00E4	a	.	n	y
Ä	Ä		U+00C4	A	.	n	y
å *å*	å		U+00E5	a	.	n	y
ã	ã		U+00E3	a	.	n	y
ắ	ắ		U+1EAF	a	.	n	y
ǟ	ǟ		U+01DF	a	.	n	y
ā́	&few-a-long-accent;		U+F400	a	.	n	y
	&few-a-ouvert-long;		U+F401	a	.	n	y
	&few-a-ouvert-long-accent;		U+F402	a	.	n	y
ǻ	&few-a-rond-accent;		TBD	a	.	n	y
	&few-a-rond-long;		TBD	a	.	n	y
	&few-a-rond-long-accent;		TBD	a	.	n	y
å̃	&few-a-rond-tilde;		TBD	a	.	n	y
å̃́	&few-a-rond-tilde-accent;		TBD	a	.	n	y
ã́	&few-a-tilde-accent;		TBD	a	.	n	y
	&few-a-tilde-long;		TBD	a	.	n	y
	&few-a-tilde-long-accent;		TBD	a	.	n	y
ä́	&few-a-trema-accent;		TBD	a	.	n	y

	&few-a-trema-long-accent;		TBD	a	.	n	y
ã̈	&few-a-trema-tilde;		TBD	a	.	n	y
	&few-a-trema-tilde-accent;		TBD	a	.	n	y
Æ	Æ		U+00C6	AE	.	n	y
æ		æ	U+00E6	ae	.	n	y
	&few-aelig-circonflexe;		TBD	ae	.	n	y
ĕ *ĕ*	ĕ		U+0115	e	.	n	y
ē *ē*	ē		U+0113	e	.	n	y
é		é	U+00E9	e	.	n	y
è		è	U+00E8	e	.	n	y
ê		ê	U+00EA	e	.	n	y
ë		ë	U+00EB	e	.	n	y
e̊	&few-e-rond;		TBD	e	.	n	y
ẹ	ẹ		U+1EB9	e	.	n	y
ẽ	ẽ		U+1EBD	e	.	n	y
ĕ́	&few-e-bref-accent;		TBD	e	.	n	y
ẹ́	&few-e-ferme-accent;		TBD	e	.	n	y
ẹ̆	&few-e-ferme-bref;		TBD	e	.	n	y
ẹ̆́	&few-e-ferme-bref-accent;		TBD	e	.	n	y
ẹ̄	&few-e-ferme-long;		TBD	e	.	n	y
ẹ̄́	&few-e-ferme-long-accent;		TBD	e	.	n	y
ẹ̃	&few-e-ferme-tilde;		TBD	e	.	n	y
	&few-e-ferme-tilde-accent;		TBD	e	.	n	y
ḗ	&few-e-long-accent;		TBD	e	.	n	y
ę	&few-e-ouvert;		TBD	e	.	n	y
ę́	&few-e-ouvert-accent;		TBD	e	.	n	y
ę̆	&few-e-ouvert-bref;		TBD	e	.	n	y
ę̆́	&few-e-ouvert-bref-accent;		TBD	e	.	n	y

ę̄	&few-e-ouvert-long;		TBD	e	.	n	y
ę̄́	&few-e-ouvert-long-accent;		TBD	e	.	n	y
ę̃	&few-e-ouvert-tilde;		TBD	e	.	n	y
	&few-e-ouvert-tilde-accent;		TBD	e	.	n	y
	&few-e-trema-circonflexe;		TBD	e	.	n	y
ẽ́	&few-e-tilde-accent;		TBD	e	.	n	y
ə	ə		U+0259	e	.	n	y
ə́	&few-schwa-accent;		TBD	e	.	n	y
ə̃	&few-schwa-tilde;		TBD	e	.	n	y
ə̃́	&few-schwa-tilde-accent;		TBD	e	.	n	y
	&few-schwa-bref;		TBD	e	.	n	y
	&few-schwa-bref-accent;		TBD	e	.	n	y
	&few-schwa-long;		TBD	e	.	n	y
	&few-schwa-long-accent;		TBD	e	.	n	y
ĭ ĭ	ĭ		U+012D	i	.	n	y
ī ī	ī		U+012B	i	.	n	y
í í	í		U+00ED	i	.	n	y
ì	ì		U+00EC	i	.	n	y
î		î	U+00EE	i	.	n	y
ï		ï	U+00EF	i	.	n	y
ĩ	ĩ		U+0129	i	.	n	y
ị	ị		U+1ECB	i	.	n	y
ĭ́	&few-i-bref-accent;		TBD	i	.	n	y
	&few-i-ferme-accent;		TBD	i	.	n	y
ī́	&few-i-long-accent;		TBD	i	.	n	y
į	&few-i-ouvert;		TBD	i	.	n	y
į́	&few-i-ouvert-accent;		TBD	i	.	n	y
	&few-i-ouvert-long;		TBD	i	.	n	y

	&few-i-ouvert-long-accent;		TBD	i	.	n	y
ĩ́	&few-i-tilde-accent;		TBD	i	.	n	y
	&few-i-semi-voyelle;		TBD	i	.	n	y
ŏ	ŏ		U+014F	o	.	n	y
ŏ *ŏ*	ŏ		U+014F	o	.	n	y
ō *ō*	ō		U+014D	o	.	n	y
ó *ó*	ó		U+00F3	o	.	n	y
ò	ò		U+00F2	o	.	n	y
ô		ô	U+00F4	o	.	n	y
ö		ö	U+00F6	o	.	n	y
	&few-o-long-circonflexe;		TBD	o	.	n	y
ǫ	&few-o-ouvert;		TBD	o	.	n	y
ọ	ọ		U+1ECD	o	.	n	y
õ	õ		U+00F5	o	.	n	y
ŏ́	&few-o-bref-accent;		TBD	o	.	n	y
ọ́	&few-o-ferme-accent;		TBD	o	.	n	y
ọ̆	&few-o-ferme-bref;		TBD	o	.	n	y
ọ̆́	&few-o-ferme-bref-accent;		TBD	o	.	n	y
ọ̄	&few-o-ferme-long;		TBD	o	.	n	y
ọ̄́	&few-o-ferme-long-accent;		TBD	o	.	n	y
	&few-o-ferme-tilde;		TBD	o	.	n	y
	&few-o-ferme-tilde-accent;		TBD	o	.	n	y
ṓ	&few-o-long-accent;		TBD	o	.	n	y
ǫ́	&few-o-ouvert-accent;		TBD	o	.	n	y
ǭ	&few-o-ouvert-long;		TBD	o	.	n	y
ǭ́	&few-o-ouvert-long-accent;		TBD	o	.	n	y
ǫ̆	&few-o-ouvert-bref;		TBD	o	.	n	y
ǫ̆́	&few-o-ouvert-bref-accent;		TBD	o	.	n	y
	&few-o-ouvert-tilde;		TBD	o	.	n	y

	&few-o-ouvert-tilde-accent;		TBD	o	.	n	y
ǒ̃	&few-o-tilde-accent;		TBD	o	.	n	y
ø	ø		U+00F8	o	.	n	y
	&few-o-slash-circonflexe;		TBD	o	.	n	y
Œ	Œ		U+0152	OE	.	n	y
œ		œ	U+0153	oe	.	n	y
	&few-oelig-accent;		TBD	oe	.	n	y
	&few-oelig-bref;		TBD	oe	.	n	y
	&few-oelig-bref-accent;		TBD	oe	.	n	y
œ̣	&few-oelig-ferme;		TBD	oe	.	n	y
œ̣́	&few-oelig-ferme-accent;		TBD	oe	.	n	y
œ̣̄	&few-oelig-ferme-long;		TBD	oe	.	n	y
œ̣̄́	&few-oelig-ferme-long-accent;		TBD	oe	.	n	y
œ̣̆	&few-oelig-ferme-bref;		TBD	oe	.	n	y
œ̣̆́	&few-oelig-ferme-bref-accent;		TBD	oe	.	n	y
	&few-oelig-ferme-tilde;		TBD	oe	.	n	y
	&few-oelig-ferme-tilde-accent;		TBD	oe	.	n	y
	&few-oelig-long;		TBD	oe	.	n	y
	&few-oelig-long-accent;		TBD	oe	.	n	y
œ̨	&few-oelig-ouvert;		TBD	oe	.	n	y
œ̨́	&few-oelig-ouvert-accent;		TBD	oe	.	n	y
œ̨̄	&few-oelig-ouvert-long;		TBD	oe	.	n	y
œ̨̄́	&few-oelig-ouvert-long-accent;		TBD	oe	.	n	y
œ̨̆	&few-oelig-ouvert-bref;		TBD	oe	.	n	y
œ̨̆́	&few-oelig-ouvert-bref-accent;		TBD	oe	.	n	y
	&few-oelig-ouvert-tilde;		TBD	oe	.	n	y
	&few-oelig-ouvert-tilde-accent;		TBD	oe	.	n	y
œ̃	&few-oelig-tilde;		TBD	oe	.	n	y
œ̃́	&few-oelig-tilde-accent;		TBD	oe	.	n	y

ŭ	ŭ		U+016D	u	.	n	y
ū	ū		U+016B	u	.	n	y
	&few-u-long-bref;		TBD	u	.	n	y
ú	ú		U+00FA	u	.	n	y
ù		ù	U+00F9	u	.	n	y
û		û	U+00FB	u	.	n	y
ü		ü	U+00FC	u	.	n	y
Ü	Ü		U+00DC	U	.	n	y
ǘ	ǘ		U+01D8	u	.	n	y
ǖ	ǖ		U+01D6	u	.	n	y
ü̆	&few-u-trema-bref;		TBD	u	.	n	y
	&few-u-trema-bref-accent;		TBD	u	.	n	y
	&few-u-trema-long-accent;		TBD	u	.	n	y
ü̜	&few-u-trema-ouvert;		TBD	u	.	n	y
	&few-u-trema-ouvert-accent;		TBD	u	.	n	y
	&few-u-trema-ouvert-bref;		TBD	u	.	n	y
	&few-u-trema-ouvert-bref-accent;		TBD	u	.	n	y
	&few-u-trema-ouvert-long;		TBD	u	.	n	y
	&few-u-trema-ouvert-long-accent;		TBD	u	.	n	y
ü̃	&few-u-trema-tilde;		TBD	u	.	n	y
ü̃́	&few-u-trema-tilde-accent;		TBD	u	.	n	y
u̯	&few-u-semi-voyelle;		TBD	u	.	n	y
u	&few-u-boucle;		TBD	u	.	n	y
ú	&few-u-boucle-accent;		TBD	u	.	n	y
ŭ	&few-u-boucle-bref;		TBD	u	.	n	y
ŭ́	&few-u-boucle-bref-accent;		TBD	u	.	n	y
	&few-u-boucle-ferme;		TBD	u	.	n	y
ū	&few-u-boucle-long;		TBD	u	.	n	y

ṵ̄́	&few-u-boucle-long-accent;		TBD	u	.	n	y
ṵ	&few-u-boucle-ouvert;		TBD	u	.	n	y
ṵ́	&few-u-boucle-ouvert-accent;		TBD	u	.	n	y
ṵ̆	&few-u-boucle-ouvert-bref;		TBD	u	.	n	y
ṵ̆́	&few-u-boucle-ouvert-bref-accent;		TBD	u	.	n	y
ṵ̄	&few-u-boucle-ouvert-long;		TBD	u	.	n	y
ṵ̄́	&few-u-boucle-ouvert-long-accent;		TBD	u	.	n	y
	&few-u-boucle-rond;		TBD	u	.	n	y
ũ	&few-u-boucle-tilde;		TBD	u	.	n	y
ṹ	&few-u-boucle-tilde-accent;		TBD	u	.	n	y
ƀ	ƀ		U+0180	b	.	n	y
ç		ç	U+00E7	c	.	n	y
č	č		U+010D	c	.	n	y
đ	đ		U+0111	d	.	n	y
ḍ	ḍ		U+1E0D	d	.	n	y
ḏ	ḏ		U+1E0F	d	.	n	y
ǧ	ǧ		U+01E7	g	.	n	y
ġ	ġ		U+0121	g	.	n	y
Ġ	Ġ		U+0120	G	.	n	y
ḫ	ḫ		U+1E2B	h	.	n	y
Ḫ	Ḫ		U+1E2A	H	.	n	y
ḥ	ḥ		U+1E25	h	.	n	y
ʿh	&few-h-apostrophe;		TBD	h	.	n	y
ḳ	ḳ		U+1E33	k	.	n	y
ł	ł		U+0142	l	.	n	y
ɫ ɫ	ɫ		U+026B	l	.	n	y
ñ ñ	ñ		U+00F1	n	.	n	y
ɳ	ɳ		U+0273	n	.	n	y
ś	ś		U+015B	s	.	n	y

þ	þ		U+00FE	th	.	n	y
ṛ	ṛ		U+1E5B	r	.	n	y
ř	ř		U+0159	r	.	n	y
ṣ	ṣ		U+1E63	s	.	n	y
š *š*	š		U+0161	s	.	n	y
ş	ş		U+015F	s	.	n	y
ş̥	&few-s-rond;		TBD	s	.	n	y
ß	ß		U+00DF	ss	.	n	n
ṭ	ṭ		U+1E6D	t	.	n	y
ṯ	ṯ		U+1E6F	t	.	n	y
ţ	ţ		U+0163	t	.	n	y
ẅ	ẅ		U+1E85	w	.	n	y
ÿ	ÿ		U+00FF	y	.	n	y
ŷ	ŷ		U+0177	y	.	n	y
ỹ	ỹ		U+1EF9	y	.	n	y
ý	ý		U+00FD	y	.	n	y
ȳ	ȳ		U+0233	y	.	n	y
ẓ	ẓ		U+1E93	z	.	n	y
ż	ż		U+017C	z	.	n	y
ź	ź		U+017A	z	.	n	y
ž *ž*	ž		U+017E	z	.	n	y
ʾ	ʾ		U+02BE	'	.	n	y
ʿ	ʿ		U+02BF	'	.	n	y
Θ	Θ		U+0398	th	.	n	y
	&few-hyphen-accent;		TBD	?	.	n	n
>	>		U+003E	>	|	n	n
<	<		U+003C	<	|	n	n

This character, different from &few-o-ouvert;, is present only in etymons

ǫ	ǫ		U+01EB	o	.	n	n

In addition to these phonetic characters, other existing Unicode characters can be used as-is, i.e. typed using standard UTF-8 coding. The general rules are

<table>
<tr><td>ascii</td><td></td><td>ascii</td><td>U+...</td><td></td><td></td><td></td><td></td></tr>
<tr><td>Unicode</td><td></td><td>Unicode</td><td>U+...</td><td></td><td></td><td></td><td></td></tr>
</table>

Among existing Unicode characters, some are of special importance; the UTF-8 values used to store them must be checked explicitly against the expected values provided here below

<table>
<tr><td>Comma</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>,</td><td></td><td>,</td><td>U+002C</td><td>,</td><td>|</td><td>n</td><td>n</td></tr>
<tr><td>Quotation</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>‚</td><td></td><td>‚</td><td>U+201A</td><td>'</td><td>|</td><td>n</td><td>n</td></tr>
<tr><td>‘</td><td></td><td>‘</td><td>U+2018</td><td>'</td><td>|</td><td>n</td><td>n</td></tr>
<tr><td>’</td><td></td><td>’</td><td>U+2019</td><td>'</td><td>|</td><td>n</td><td>n</td></tr>
<tr><td>„</td><td></td><td>„</td><td>U+201E</td><td>"</td><td>|</td><td>n</td><td>n</td></tr>
<tr><td>“</td><td></td><td>“</td><td>U+201C</td><td>"</td><td>|</td><td>n</td><td>n</td></tr>
<tr><td>”</td><td></td><td>”</td><td>U+201D</td><td>"</td><td>|</td><td>n</td><td>n</td></tr>
<tr><td>Others</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>→</td><td></td><td>→</td><td>U+2192</td><td>?</td><td>|</td><td>n</td><td>n</td></tr>
<tr><td>∅</td><td></td><td>∅</td><td>U+2205</td><td>?</td><td>|</td><td>n</td><td>n</td></tr>
<tr><td>⌜</td><td></td><td>⌜</td><td>U+231C</td><td>?</td><td>|</td><td>n</td><td>n</td></tr>
<tr><td>⌝</td><td></td><td>⌝</td><td>U+231D</td><td>?</td><td>|</td><td>n</td><td>n</td></tr>
<tr><td>-</td><td></td><td>-</td><td>U+002D</td><td>-</td><td>|</td><td>n</td><td>n</td></tr>
<tr><td>–</td><td></td><td>–</td><td>U+2013</td><td>-</td><td>|</td><td>n</td><td>n</td></tr>
<tr><td>—</td><td></td><td>—</td><td>U+2014</td><td>-</td><td>|</td><td>n</td><td>n</td></tr>
</table>

Others common characters are mentioned here below because they're words delimiters

<table>
<tr><td></td><td></td><td></td><td>U+0020</td><td></td><td>|</td><td>n</td><td>n</td></tr>
<tr><td>!</td><td></td><td>!</td><td>U+0021</td><td>!</td><td>|</td><td>n</td><td>n</td></tr>
<tr><td>(</td><td></td><td>(</td><td>U+0028</td><td>(</td><td>|</td><td>n</td><td>n</td></tr>
</table>

)		)	U+0029	)	|	n	n
*		*	U+002A	*	|	n	n
+		+	U+002B	+	|	n	n
.		.	U+002E	.	|	n	n
:		:	U+003A	:	|	n	n
;		;	U+003B	;	|	n	n
=		=	U+003D	=	|	n	n
?		?	U+003F	?	|	n	n
[		[	U+005B	[	|	n	n
]		]	U+005D	]	|	n	n
/		/	U+002F	/	|	n	n
†		†	U+2020	?	|	n	n
…		…	U+2026	...	|	n	n
'		'	U+0027	'	|	n	n

The characters of the Greek alphabet (including all diacritics) follow the general rules, e.g.

α		α	U+03B1	a	.	y	n
β		β	U+03B2	b	.	y	y
γ		γ	U+03B3	g	.	y	y
δ		δ	U+03B4	d	.	y	y
ε		ε	U+03B5	e	.	y	n
ζ		ζ	U+03B6	z	.	y	n
η		η	U+03B7	e	.	y	n
θ		θ	U+03B8	th	.	y	y
ι		ι	U+03B9	i	.	y	n
κ		κ	U+03BA	k	.	y	n
λ		λ	U+03BB	l	.	y	n
μ		μ	U+03BC	m	.	y	n

ν		ν	U+03BD	n	.	y	n
ξ		ξ	U+03BE	x	.	y	n
ο		ο	U+03BF	o	.	y	n
π		π	U+03C0	p	.	y	n
ρ		ρ	U+03C1	r	.	y	n
ς		ς	U+03C2	s	.	y	n
σ		σ	U+03C3	s	.	y	n
τ		τ	U+03C4	t	.	y	n
υ		υ	U+03C5	u	.	y	n
φ		φ	U+03C6	ph	.	y	y
χ		χ	U+03C7	ch	.	y	y
ψ		ψ	U+03C8	ps	.	y	n
Ω		ω	U+03C9	o	.	y	n

The accentuated Greek characters can be downloaded from
www.unicode.org/charts/PDF/U0370.pdf
www.unicode.org/charts/PDF/U1F00.pdf

Ά		Ά	U+0386	?	.	y	n
·		·	U+0387	?	.	y	n
Έ		Έ	U+0388	?	.	y	n
Ή		Ή	U+0389	?	.	y	n
Ί		Ί	U+038A	?	.	y	n
Ό		Ό	U+038C	?	.	y	n
Ύ		Ύ	U+038E	?	.	y	n
Ώ		Ώ	U+038F	?	.	y	n
Α		Α	U+0391	?	.	y	n
Β		Β	U+0392	?	.	y	n
Γ		Γ	U+0393	?	.	y	n
Δ		Δ	U+0394	?	.	y	n

Ε		Ε	U+0395	?	.	y	n
Ζ		Ζ	U+0396	?	.	y	n
Η		Η	U+0397	?	.	y	n
Θ	*cf. here above*						
Ι		Ι	U+0399	?	.	y	n
Κ		Κ	U+039A	?	.	y	n
Λ		Λ	U+039B	?	.	y	n
Μ		Μ	U+039C	?	.	y	n
Ν		Ν	U+039D	?	.	y	n
Ξ		Ξ	U+039E	?	.	y	n
Ο		Ο	U+039F	?	.	y	n
Π		Π	U+03A0	?	.	y	n
Ρ		Ρ	U+03A1	?	.	y	n
Σ		Σ	U+03A3	?	.	y	n
Τ		Τ	U+03A4	?	.	y	n
Υ		Υ	U+03A5	?	.	y	n
Φ		Φ	U+03A6	?	.	y	n
Χ		Χ	U+03A7	?	.	y	n
Ψ		Ψ	U+03A8	?	.	y	n
Ω		Ω	U+03A9	?	.	y	n
Ϊ		Ϊ	U+03AA	?	.	y	n
Ϋ		Ϋ	U+03AB	?	.	y	n
ά		ά	U+03AC	?	.	y	n
έ		έ	U+03AD	?	.	y	n
ή		ή	U+03AE	?	.	y	n
ί		ί	U+03AF	?	.	y	n
ΰ		ΰ	U+03B0	?	.	y	n
ό		ό	U+03CC	?	.	y	n
ύ		ύ	U+03CD	?	.	y	n
ώ		ώ	U+03CE	?	.	y	n

ἀ		ἀ	U+1F00	?	.	y	n
ἁ		ἁ	U+1F01	?	.	y	n
ἂ		ἂ	U+1F02	?	.	y	n
ἃ		ἃ	U+1F03	?	.	y	n
ἄ		ἄ	U+1F04	?	.	y	n
ἅ		ἅ	U+1F05	?	.	y	n
ἆ		ἆ	U+1F06	?	.	y	n
ἇ		ἇ	U+1F07	?	.	y	n
Ἀ		Ἀ	U+1F08	?	.	y	n
Ἁ		Ἁ	U+1F09	?	.	y	n
Ἂ		Ἂ	U+1F0A	?	.	y	n
Ἃ		Ἃ	U+1F0B	?	.	y	n
Ἄ		Ἄ	U+1F0C	?	.	y	n
Ἅ		Ἅ	U+1F0D	?	.	y	n
Ἆ		Ἆ	U+1F0E	?	.	y	n
Ἇ		Ἇ	U+1F0F	?	.	y	n
ἐ		ἐ	U+1F10	?	.	y	n
ἑ		ἑ	U+1F11	?	.	y	n
ἒ		ἒ	U+1F12	?	.	y	n
ἓ		ἓ	U+1F13	?	.	y	n
ἔ		ἔ	U+1F14	?	.	y	n
ἕ		ἕ	U+1F15	?	.	y	n
Ἐ		Ἐ	U+1F18	?	.	y	n
Ἑ		Ἑ	U+1F19	?	.	y	n
Ἒ		Ἒ	U+1F1A	?	.	y	n
Ἓ		Ἓ	U+1F1B	?	.	y	n
Ἔ		Ἔ	U+1F1C	?	.	y	n
Ἕ		Ἕ	U+1F1D	?	.	y	n
ἠ		ἠ	U+1F20	?	.	y	n
ἡ		ἡ	U+1F21	?	.	y	n

ἢ		ἢ	U+1F22	?	.	y	n
ἣ		ἣ	U+1F23	?	.	y	n
ἤ		ἤ	U+1F24	?	.	y	n
ἥ		ἥ	U+1F25	?	.	y	n
ἦ		ἦ	U+1F26	?	.	y	n
ἧ		ἧ	U+1F27	?	.	y	n
Ἠ		Ἠ	U+1F28	?	.	y	n
Ἡ		Ἡ	U+1F29	?	.	y	n
Ἢ		Ἢ	U+1F2A	?	.	y	n
Ἣ		Ἣ	U+1F2B	?	.	y	n
Ἤ		Ἤ	U+1F2C	?	.	y	n
Ἥ		Ἥ	U+1F2D	?	.	y	n
Ἦ		Ἦ	U+1F2E	?	.	y	n
Ἧ		Ἧ	U+1F2F	?	.	y	n
ἰ		ἰ	U+1F30	?	.	y	n
ἱ		ἱ	U+1F31	?	.	y	n
ἲ		ἲ	U+1F32	?	.	y	n
ἳ		ἳ	U+1F33	?	.	y	n
ἴ		ἴ	U+1F34	?	.	y	n
ἵ		ἵ	U+1F35	?	.	y	n
ἶ		ἶ	U+1F36	?	.	y	n
ἷ		ἷ	U+1F37	?	.	y	n
Ἰ		Ἰ	U+1F38	?	.	y	n
Ἱ		Ἱ	U+1F39	?	.	y	n
Ἲ		Ἲ	U+1F3A	?	.	y	n
Ἳ		Ἳ	U+1F3B	?	.	y	n
Ἴ		Ἴ	U+1F3C	?	.	y	n
Ἵ		Ἵ	U+1F3D	?	.	y	n
Ἶ		Ἶ	U+1F3E	?	.	y	n
Ἷ		Ἷ	U+1F3F	?	.	y	n

ὀ		ὀ	U+1F40	?	.	y	n
ὁ		ὁ	U+1F41	?	.	y	n
ὂ		ὂ	U+1F42	?	.	y	n
ὃ		ὃ	U+1F43	?	.	y	n
ὄ		ὄ	U+1F44	?	.	y	n
ὅ		ὅ	U+1F45	?	.	y	n
Ὀ		Ὀ	U+1F48	?	.	y	n
Ὁ		Ὁ	U+1F49	?	.	y	n
Ὂ		Ὂ	U+1F4A	?	.	y	n
Ὃ		Ὃ	U+1F4B	?	.	y	n
Ὄ		Ὄ	U+1F4C	?	.	y	n
Ὅ		Ὅ	U+1F4D	?	.	y	n
ὐ		ὐ	U+1F50	?	.	y	n
ὑ		ὑ	U+1F51	?	.	y	n
ὒ		ὒ	U+1F52	?	.	y	n
ὓ		ὓ	U+1F53	?	.	y	n
ὔ		ὔ	U+1F54	?	.	y	n
ὕ		ὕ	U+1F55	?	.	y	n
ὖ		ὖ	U+1F56	?	.	y	n
ὗ		ὗ	U+1F57	?	.	y	n
Ὑ		Ὑ	U+1F59	?	.	y	n
Ὓ		Ὓ	U+1F5B	?	.	y	n
Ὕ		Ὕ	U+1F5D	?	.	y	n
Ὗ		Ὗ	U+1F5F	?	.	y	n
ὠ		ὠ	U+1F60	?	.	y	n
ὡ		ὡ	U+1F61	?	.	y	n
ὢ		ὢ	U+1F62	?	.	y	n
ὣ		ὣ	U+1F63	?	.	y	n
ὤ		ὤ	U+1F64	?	.	y	n
ὥ		ὥ	U+1F65	?	.	y	n

ὦ		ὦ	U+1F66	?	.	y	n
ὧ		ὧ	U+1F67	?	.	y	n
Ὠ		Ὠ	U+1F68	?	.	y	n
Ὡ		Ὡ	U+1F69	?	.	y	n
Ὢ		Ὢ	U+1F6A	?	.	y	n
Ὣ		Ὣ	U+1F6B	?	.	y	n
Ὤ		Ὤ	U+1F6C	?	.	y	n
Ὥ		Ὥ	U+1F6D	?	.	y	n
Ὦ		Ὦ	U+1F6E	?	.	y	n
Ὧ		Ὧ	U+1F6F	?	.	y	n
ὰ		ὰ	U+1F70	?	.	y	n
ά		ά	U+1F71	?	.	y	n
ὲ		ὲ	U+1F72	?	.	y	n
έ		έ	U+1F73	?	.	y	n
ὴ		ὴ	U+1F74	?	.	y	n
ή		ή	U+1F75	?	.	y	n
ὶ		ὶ	U+1F76	?	.	y	n
ί		ί	U+1F77	?	.	y	n
ὸ		ὸ	U+1F78	?	.	y	n
ό		ό	U+1F79	?	.	y	n
ὺ		ὺ	U+1F7A	?	.	y	n
ύ		ύ	U+1F7B	?	.	y	n
ὼ		ὼ	U+1F7C	?	.	y	n
ώ		ώ	U+1F7D	?	.	y	n
ᾀ		ᾀ	U+1F80	?	.	y	n
ᾁ		ᾁ	U+1F81	?	.	y	n
ᾂ		ᾂ	U+1F82	?	.	y	n
ᾃ		ᾃ	U+1F83	?	.	y	n
ᾄ		ᾄ	U+1F84	?	.	y	n
ᾅ		ᾅ	U+1F85	?	.	y	n

ᾆ		ᾆ	U+1F86	?	.	y	n
ᾇ		ᾇ	U+1F87	?	.	y	n
Ἀι		Ἀι	U+1F88	?	.	y	n
Ἁι		Ἁι	U+1F89	?	.	y	n
Ἂι		Ἂι	U+1F8A	?	.	y	n
Ἃι		Ἃι	U+1F8B	?	.	y	n
Ἄι		Ἄι	U+1F8C	?	.	y	n
Ἅι		Ἅι	U+1F8D	?	.	y	n
Ἆι		Ἆι	U+1F8E	?	.	y	n
Ἇι		Ἇι	U+1F8F	?	.	y	n
ᾐ		ᾐ	U+1F90	?	.	y	n
ᾑ		ᾑ	U+1F91	?	.	y	n
ᾒ		ᾒ	U+1F92	?	.	y	n
ᾓ		ᾓ	U+1F93	?	.	y	n
ᾔ		ᾔ	U+1F94	?	.	y	n
ᾕ		ᾕ	U+1F95	?	.	y	n
ᾖ		ᾖ	U+1F96	?	.	y	n
ᾗ		ᾗ	U+1F97	?	.	y	n
Ἠι		Ἠι	U+1F98	?	.	y	n
Ἡι		Ἡι	U+1F99	?	.	y	n
Ἢι		Ἢι	U+1F9A	?	.	y	n
Ἣι		Ἣι	U+1F9B	?	.	y	n
Ἤι		Ἤι	U+1F9C	?	.	y	n
Ἥι		Ἥι	U+1F9D	?	.	y	n
Ἦι		Ἦι	U+1F9E	?	.	y	n
Ἧι		Ἧι	U+1F9F	?	.	y	n
ᾠ		ᾠ	U+1FA0	?	.	y	n
ᾡ		ᾡ	U+1FA1	?	.	y	n
ᾢ		ᾢ	U+1FA2	?	.	y	n
ᾣ		ᾣ	U+1FA3	?	.	y	n

ᾤ		ᾤ	U+1FA4	?	.	y	n
ᾥ		ᾥ	U+1FA5	?	.	y	n
ᾦ		ᾦ	U+1FA6	?	.	y	n
ᾧ		ᾧ	U+1FA7	?	.	y	n
ᾨ		ᾨ	U+1FA8	?	.	y	n
ᾩ		ᾩ	U+1FA9	?	.	y	n
ᾪ		ᾪ	U+1FAA	?	.	y	n
ᾫ		ᾫ	U+1FAB	?	.	y	n
ᾬ		ᾬ	U+1FAC	?	.	y	n
ᾭ		ᾭ	U+1FAD	?	.	y	n
ᾮ		ᾮ	U+1FAE	?	.	y	n
ᾯ		ᾯ	U+1FAF	?	.	y	n
ᾰ		ᾰ	U+1FB0	?	.	y	n
ᾱ		ᾱ	U+1FB1	?	.	y	n
ᾲ		ᾲ	U+1FB2	?	.	y	n
ᾳ		ᾳ	U+1FB3	?	.	y	n
ᾴ		ᾴ	U+1FB4	?	.	y	n
ᾶ		ᾶ	U+1FB6	?	.	y	n
ᾷ		ᾷ	U+1FB7	?	.	y	n
Ᾰ		Ᾰ	U+1FB8	?	.	y	n
Ᾱ		Ᾱ	U+1FB9	?	.	y	n
Ὰ		Ὰ	U+1FBA	?	.	y	n
Ά		Ά	U+1FBB	?	.	y	n
ᾼ		ᾼ	U+1FBC	?	.	y	n
᾽		᾽	U+1FBD	?	.	y	n
ι		ι	U+1FBE	?	.	y	n
			U+1FBF	?	.	y	n
῀		῀	U+1FC0	?	.	y	n
῁		῁	U+1FC1	?	.	y	n
ῂ		ῂ	U+1FC2	?	.	y	n

ῃ		ῃ	U+1FC3	?	.	y	n
ῄ		ῄ	U+1FC4	?	.	y	n
ῆ		ῆ	U+1FC6	?	.	y	n
ῇ		ῇ	U+1FC7	?	.	y	n
Ὲ		Ὲ	U+1FC8	?	.	y	n
Έ		Έ	U+1FC9	?	.	y	n
Ὴ		Ὴ	U+1FCA	?	.	y	n
Ή		Ή	U+1FCB	?	.	y	n
ῌ		ῌ	U+1FCC	?	.	y	n
῍		῍	U+1FCD	?	.	y	n
῎		῎	U+1FCE	?	.	y	n
῏		῏	U+1FCF	?	.	y	n
ῖ		ῖ	U+1FD6	?	.	y	n
Ὶ		Ὶ	U+1FDA	?	.	y	n
Ί		Ί	U+1FDB	?	.	y	n
῝		῝	U+1FDD	?	.	y	n
῞		῞	U+1FDE	?	.	y	n
῟		῟	U+1FDF	?	.	y	n
ῤ		ῤ	U+1FE4	?	.	y	n
ῥ		ῥ	U+1FE5	?	.	y	n
ῦ		ῦ	U+1FE6	?	.	y	n
Ὺ		Ὺ	U+1FEA	?	.	y	n
Ύ		Ύ	U+1FEB	?	.	y	n
Ῥ		Ῥ	U+1FEC	?	.	y	n
`		`	U+1FEF	?	.	y	n
ῲ		ῲ	U+1FF2	?	.	y	n
ῳ		ῳ	U+1FF3	?	.	y	n
ῴ		ῴ	U+1FF4	?	.	y	n
ῶ		ῶ	U+1FF6	?	.	y	n
ῷ		ῷ	U+1FF7	?	.	y	n

Ὸ		Ὸ	U+1FF8	?	.	y	n
Ό		Ό	U+1FF9	?	.	y	n
Ὼ		Ὼ	U+1FFA	?	.	y	n
Ώ		Ώ	U+1FFB	?	.	y	n
ῼ		ῼ	U+1FFC	?	.	y	n
´		´	U+1FFD	?	.	y	n
῾		῾	U+1FFE	?	.	y	n

Annexe D

FFML

D.1 Introduction

La présente annexe contient deux sections.

La première section met à disposition du lecteur de cette thèse le guide FFML, qui décrit le format de document attendu en entrée du logiciel de rétroconversion (→ 4.3).

Le schéma XML correspondant au format FFML se trouve dans le CD-ROM fourni avec cette thèse (→ G.2).

En guise d'exemple, la seconde section de cette annexe présente un exemple d'article du FEW mis en forme selon les spécifications de ce format.

D.2 Guide FFML

============================ ***draft v20101105*** ============================

Introduction to the FEW Font-style Markup Language (FFML)

Technical report, November 2010

Pascale Renders (1) (2), Cyril Briquet (2), Etienne Petitjean (2)
Pascale.Renders@ulg.ac.be, *Cyril.Briquet@acm.org*, *Etienne.Petitjean@atilf.fr*

(1) University of Liège (2) ATILF/CNRS Nancy

Abstract

FEW Font-style Markup Language (FFML) is an XML-based language that is designed to store digitized versions of articles of the Französisches Etymologisches Wörterbuch (FEW). FFML addresses the issue of describing digitized versions of FEW articles in an open way. It is the first and a very important step towards the retroconversion of the FEW, which is of high interest to the community of historical Romance linguistics.

A previous similar initiative was based on proprietary software, i.e. Microsoft Word. While such a storage scheme is obviously eye-pleasing to the human user, it is not suitable for software processing due to the proprietary and closed nature of the storage format.

FFML provides a small set of tags (total: 12 tags) to describe the formatting (6 tags) as well as the basic structure (5 tags) and provenance (1 tag) of an FEW article. It could be easily extended to include a larger number of tags. While some effort has been made towards convergence with TEI markups, full compatibility was not an objective. Instead, tag names were shortened as much as possible, in order to minimize the cost of manual encoding. A reference implementation of an FFML validator is offered (ffml-validator) and enables to check the compliance of XML documents with the schema of FFML that is defined in this document.

Keywords

FEW, XML, XML Schema, guidelines, Unicode

Contents

Section 1 - Introduction

This document is an introductory guide to the FEW Font-style Markup Language (FFML). FFML is an XML-based markup language. It is designed to store digitized versions of FEW articles in an open and standardized way.

An FFML document is basically a sequence of FEW articles. As FFML is an XML-based markup language, it can be strictly defined. Everyone can freely learn it and develop software to write or read FFML documents.

FFML is mostly independent from the digitizing process, which can rely on manual encoding, or scanning + ocerization + basic tagging, or any other digitizing mechanism. FFML is also independent from the processing and exploitation to which FFML-compliant documents can be submitted, although its design has been driven by the project of retroconversion of the FEW. Most design decisions were made by Pascale Renders in the context of her doctoral research, with inputs from Etienne Petitjean to facilitate future convergence with TEI markups.

The current version of FFML (v0.8) has been defined as an XML schema by Cyril Briquet within the scope of the DETCOL project. XML schema was selected because it is XML-based and because it is widely supported.

The rest of this document is structured as follows. Section 2 defines the structural tags of FFML. These enable to describe set of articles and then articles. Section 3 defines the formatting tags of FFML. These tags enable to describe the font-style and page layout of an article. Section 4 defines the provenance tags of FFML. Section 5 discusses issues related to character coding in UTF-8. Section 6 is a complete example of a FEW article stored as an FFML document. Section 7 discusses how to check and enforce compliance with the schema of FFML. An Appendix lists the documents associated to this introductory guide.

Section 2 - Structural Tags

An FFML document is basically a sequence of FEW articles. Structural tags describe the structure of an FEW article (article-level tags) or the structure of a sequence of FEW articles (dictionary-level tags).

Section 2.1 - Dictionary-level Structural Tags

There are 2 dictionary-level structural tags: dictionary (<few>) and article (<art>).
Both dictionary-level structural tags hold contents.
Both dictionary-level structural tags have attributes.

An FFML document basically consists of a single <few> tag. The <few> tag contains a non-empty sequence of one or more <art> tags. Information between articles are not preserved in an FFML document. Specifically, the following are not encoded into an FFML document: in volume 20, the section titles; in volumes 21-23, the concept classification; in other volumes, the heading letters.

The intent of FFML is to group, within an <few> tag, only articles that belongs to the same FEW book. This intent is enforced in the language definition. The articles can be split downstream (i.e. post-encoding) into separate files before they are retroconverted (a tool is provided to this end, split-ffml-volume).

In practice, three attributes must be defined for an <few> tag: volume, book and xmlns. The <few> tag intrinsically has two attributes: volume number (volume) and book number (book) which are strictly positive integers lower than or equal to, respectively, 26 and 3. The volume attribute can assume a special value of 26 if the article it contains are so-called *articles de la refonte de la lettre -B* (that are published online since 2006, see http://www.atilf.fr/few). The book attribute is mandatory, and assumes a default value of 1. The XML namespace of FFML is http://www.atilf.fr/few/ffml (see Section 6 Compliance). Providing it as a value of the xmlns attribute of <few> simplifies the XML syntax of the document, which is strongly recommended.

The <art> tag has two attributes, an article identifier (id), which is a non-negative integer, and the in-column index (ici), which is a strictly positive integer.

The id attribute is optional in most contexts, notably for the manual encoding of FEW articles. Article identifiers are intended to be automatically defined by a software tool, i.e. not by human operators. For single-article FFML documents only, and if required by the context, the <art> id attribute may be manually defined; the recommended value in such case is zero (0). By default, an FFML document is valid with or without id attributes defined for <art> tags. If present, the <art> id attribute will be read or ignored, given the context.

The ici attribute enables to reconstruct the order in which a set of separately-ocerized articles are ordered within a given column of an FEW book. The ici attribute is introduced (along with the pg and s attributes of the <col> tag, see Section 3.2) to guarantee that this order can be reconstructed, so that the visualization of virtual FEW pages is accurate, i.e. articles within a same column are displayed in the correct order.

The ici attribute is thus mandatory. It indicates the index of the article within its first column (and thus first page). For instance, if a given article is the first article beginning in a column (whether or not the end of a previous article is found beforehands), the article's in-column index is 1. If two - obviously short - articles are preceding a given article in a column, the article's in-column index is 3.

The skeleton of an FFML document storing three articles (e.g. for instance, belonging to FEW volume 2, book 1) looks as follows (comments are for exemplar purposes only, and never required):

```
<?xml version="1.0" encoding="UTF-8"?>
<few xmlns="http://www.atilf.fr/few/ffml" volume="6" book="1">
<art id="12345" ici="1"> ... </art> <!-- FEW 6/1, 51b, MAGOS -->
<art id="12346" ici="2"> ... </art> <!-- FEW 6/1, 51b, MAGUS -->
<art id="12347" ici="1"> ... </art>
</few>
```

Remark:
As will be discussed in Section 4, the encoding of an FFML document must be UTF-8.
End of remark.

The structure of an <art> tag's contents is defined in the next section.

Section 2.2 - Article-level Tags

There are 3 article-level structural tags: header (<h>), paragraph (<p>) and notes (<notes >).
All article-level structural tags hold contents. No article-level structural tag has attributes.

The contents of an <art> tag must comply with this structure:
- header (required)
- arbitrary number of paragraphs, columns and line breaks (optional)
- notes (optional)

Columns and line breaks (cf. infra, Section 3.2 Page Layout Tags) can appear everywhere. However, no column can appear in the header.

The skeleton of an article thus looks as follows:

```
<art id="24" ici="1">
        <col pg="982" s="b"/>
        <h>this is the article's header followed by a line break</h>

        <p>arbitrary number of paragraphs,<lb />
        columns or line breaks</p>

        <notes><p>finally, there may be notes</p></notes>
</art>
```

Headers and paragraphs are composed of text and of an arbitrary number of formatting tags (cf. infra, Section 3 Formatting Tags).

The notes section is comprised of at least one paragraph.

Section 3 - Formatting Tags

Formatting tags describe the formatting of an FEW article, as it can be found in the paper version of the article. There are two types of formatting tags: font-style tags and page layout tags.

Section 3.1 - Font-style Tags

There are 4 font-style tags: bold (), italic (<i>), small caps (<sc>) and exponent (<e>).
All font-style tags hold contents. No font-style tag has attributes.

Exponent tags never contain any other tag.

Bold and italic tags never contain any other tag, except page layout tags.
There is at least one bold tag in an <h> tag (this is a rule of the FEW).

e.g. <h> this is an example text chunk with two bold words</h> -- OK
e.g. <h> this is an example text chunk with no bold word</h> -- illegal

Small caps tags never contain any other tag, except page layout tags and exponent tags. A small caps tag contains at most one exponent tag.

e.g. <sc> this is an example text chunk with an <e>exponent</e></sc> -- OK
e.g. <sc> this is an example text chunk with <e>two</e> <e>exponent</e></sc> -- illegal

Section 3.2 - Page Layout Tags

There are 2 page layout tags: line break (<lb>) and column (<col>).
Both page layout tags are empty.

Line break tags have no attribute.

Column tags have two attributes: page number (pg) and column side (s).

The purpose of page layout tags is to provide layout information for articles. A column tag is inserted within articles each time a new column is encountered. A column tag is also inserted immediately at the beginning of each article, so that the beginning of the article can be situated within the FEW. A line break is inserted within articles each time a new line break is encountered. It is recommended to not insert line break tags between paragraphs, as these will most likely be automatically removed by the retroconversion software.

The pg attribute of a column tag is a strictly positive integer that designates the number of the page holding the contents provided immediately after the tag.

The s attribute of a column tag indicates the side of the column on the page. It is a string that can take one of two conventionally chosen values: a or b, which designate, respectively, the left and the right column of the page including the contents provided immediately after the tag. Given the definition of the column tag, *at least* two column tags will often be associated to each page of the FEW. One column tag marks the beginning of the left column of a page. One column tag marks the beginning of the right column of a page. It is recommended to add column tags within paragraphs, rather than in-between paragraphs.

Section 4 - Provenance Tags

Provenance tags describe the sequence of operations applied to an FEW article, as well as associated timestamps. Provenance tags are always optional, but their use is encouraged.

There is 1 provenance tag (<provenance>).

Provenance tags have three attributes: operation (op), type (type) time stamp (timestamp). The operation attribute is the label of the retroconversion algorithm, tool or manual encoding operation applied to update the FEW article. The type attribute can be either algo, manual or tool. The time stamp attribute follows a standard, widespread representation of time. It is expressed as the number of milliseconds since the Epoch (January 1, 1970, midnight, GMT). Online tools (e.g. http://www.epochconverter.com/) can be used to compute such time stamps. If relying on values computed by such online tools, one must make sure to add three zeros to express values in milliseconds. Provenance tags are located immediately prior to the closing article tag.

Example 1:

```
...
<provenance op="OCR Software XYZ" type="tool" timestamp="1239947167008"/>
<provenance op="ffml-volume-splitter" type="tool" timestamp="1239977167008"/>
<provenance op="identity" type="algo" timestamp="1239977179814"/>
<provenance op="tag-lang-etymon" type="algo" timestamp="1239977179818"/>
</art>
</few>
```

Example 2:

```
...
<provenance op="Isabelle Clément" type="manual" timestamp="1276761600000"/>
</art>
</few>
```

Provenance tags are intended (Example 1) to be automatically inserted by retroconversion algorithms or tools (e.g. FFML volume splitter tool or OCR software). Provenance tags of the manual type are also intended (Example 2) to document the manual encoding of the article, if so desired.

Section 5 - Character Encoding

Unicode is used to represent FEW characters , i.e. all characters that may appear in the FEW. The 8-bit UCS/Unicode Transformation Format (UTF-8) is selected, as can be expected, as the character encoding format for the digitized articles of the FEW. The rationale is the standard, open and widespread nature of UTF-8. However, several issues need to be addressed.

Many accentuated characters appear in the FEW. Among those, many are not in widespread use. Moreover, many of the FEW characters are not even (yet) included in Unicode/UTF-8. Two issues arise from this couple of observations. The first issue is meaningful visualization; the second issue is consistent storage.

Meaningful visualization of FEW characters consists of visualizing all FEW characters as nice-looking, meaningful glyphs, instead of weird-looking hexadecimal values, e.g. A vs. U+0065. Meaningful visualization is hard to achieve. Typically, additional fonts have to be

installed on the operating system of the computer from which visualization takes place. Moreover, it may be hard to find fonts including glyphs to represent nonstandardized characters or standardized characters that are not in widespread use. An effort orthogonal to the definition of FFML is the design of an FEW font that supports all FEW characters. Such an FEW font should be made available as soon as possible, because there is no other practical way to meaningfully visualize digitized FEW articles. Meaningful visualization of FEW characters is an open issue, though initial research efforts have been undertaken at ATILF. Consistent storage of FEW characters is not straightforward, but it is achievable. The key components to our proposed approach are XML Entities and UTF-8 private use area. We first introduce the following classification of FEW characters as a basis to our approach:

- ASCII characters
- standardized-but-rare characters
- standardized-and-in-widespread-use characters
- nonstandardized characters

A complete list of non-ASCII FEW characters is provided in the FEW character map (see Appendix). Though ASCII characters are not listed in the FEW character map, they are implicitly considered as FEW characters and thus implicitly considered as listed in the FEW character map.

We propose to encode the four aforementioned classes of FEW characters according to the following guidelines:

Firstly and obviously, we propose that ASCII characters be typed as-is.

Secondly, we propose that standardized-but-rare characters - such as U+0101 (i.e. letter “a”, topped by an horizontal bar) - be typed as numeric character references, e.g. “ā” . They will be automatically converted to their UTF-8 value by the XML parser processing FFML documents.

Thirdly, we propose that standardized-and-in-widespread-use characters - such as “à” - be directly typed as-is, e.g. “à”. Originally, these characters were intended to be typed as numeric character references too. However, as keyboard and font support are likely to be available, it was later decided to type them directly so as to facilitate the manual encoding of an FEW article. They will not be converted, as they will directly be stored using the correct UTF-8 value.

Fourthly, we propose that nonstandardized characters - such as the letter “a”, topped by an horizontal bar and an accent - be typed as character entity references, e.g. “&few-a-long-accent;”. Any processing, including validation, of FFML documents requires that XML Entities be defined for the nonstandardized characters, so that they can be converted by the XML parser. A character map of FEW characters to Unicode values is provided to this end (see Appendix). The FEW character map is embedded with the reference implementation of the FFML validator (see Section 6 Compliance). The (output) Unicode values corresponding to the (input) nonstandardized characters are arbitrarily drawn from Unicode basic multilingual plane's private use area. The FEW character map is likely to change when the design of the much-expected FEW font is finalized. For instance, these Unicode values will probably move out of the private use area.

Remarks:
1/ Users who want to write FFML documents need not be concerned with the output values of the FEW character map. FFML document writers need only to know - and use - the input values, i.e. XML Entities. Indeed, until the FEW font is available, the output Unicode values selected for the nonstandardized characters can be totally arbitrary... no one will be able to visualize them meaningfully anyways. Updating the output Unicode values of the FEW character map has little impact on the FEW retroconversion project: As a layer of indirection, the FEW character map separates the issues of nonstandardized characters encoding from the writing of an FFML document. Users who want to implement their own software tools to read FFML documents may use the most recent version of the FEW character map as necessary.

2/ It must be remarked that non-ascii characters are expected only in text chunks, and never in FFML tag names or attributes.

3/ In the context of this guide to FFML, the fourth to the eigth columns of the FEW character map are provided for informational purposes. None of them are needed to write FFML documents.
End of remarks.

In a nutshell, determining how to type a glyph belonging to the FEW characters, on the basis of the FEW character map (see Appendix), is straightforward. The procedure can be summarized as follows:

- first seek the glyph in the first column of the FEW character map (Glyph),
- then look right to the second or third column of the map (FFML & keyword encoding).

Users who want to edit an FFML using an XML editor might want to temporarily add the following line at the beginning of the FFML file:
<!DOCTYPE few SYSTEM "few-char-map-entity.dtd">
The few-char-map-entity.dtd file must be located in the directory where the FFML file is stored (or the path in the aforementioned XML doctype directive must be adjusted). This DTD file can be obtained with a DTD generation tool that is also provided (few-char-map-entity-dump). The reason why the DTD should be generated using the provided tool is that the character map is likely to evolve, and it is of paramount importance to rely on its latest version. It is important to note that the aforementioned XML doctype directive must be removed from the FFML file before it can be processed by the retroconversion software.

Section 6 - A Complete Example

Reference: FEW 2, 982b, COMPLETUS

```
<?xml version="1.0" encoding="UTF-8"?>
<few xmlns="http://www.atilf.fr/few/ffml" volume="2" book="2">
<art id="0" ici="1">

<col pg="982" s="b"/>
<h><b>completus</b> vollständig;<lb/>
vollkommen.</h><lb/>
```

<p>I. 1. a. Vollständig. — Mfr. nfr. <i>complet</i> „à<lb/>
quoi il ne manque aucune des parties nécessaires“<lb/>
(seit ca. 1300, Monstr; Rhlitt 6, 464), saint. St-<lb/>
Seurin <i>compiet</i>, Minot <i>conpiet</i>, npr. <i>coumplèt</i>. —<lb/>
Übertragen. Nfr. <i>complet</i> „(pop.) tout à fait ivre“<lb/>
(seit Flick 1802). — Ablt. Nfr. <i>se compléter</i> „ache-<lb/>
ver de s'enivrer“ (pop., seit Littré 1863).</p>
<p>Ablt. — Afr. mfr. <i>completement</i> „d'une ma-<lb/>
nière complète“ (13. jh.—1503, TL; Gdf; RF 32,<lb/>
83), nfr. <i>complètement</i> (La Fontaine 1695, dann<lb/>
seit Trév 1752)<e>1</e>), sav. <i>complletament</i>, lang. <i>coum-<lb/>
pletomen</i> M. — Nfr. <i>compléter</i> „rendre complet“<lb/>
(seit Trév 1725), saint. <i>compiéter</i>, sav. <i>complletâ</i>,<lb/>
npr. <i>coumpleta</i>, mars. <i>coumpletar</i> A. — Nfr. <i>com-<lb/>
plètement</i> „action de mettre au complet“ (seit 1750,<lb/>
text in Fér 1787). — Npr. <i>coumpletage</i>, mars.<lb/>
<i>coumpletàgi</i> M. — Nfr. <i>compléteur</i> „celui qui com-<lb/>
plète“ (seit Lar 1922, auch adj.).</p>
<p>Nfr. <i>décompléter</i> „rendre incomplet“ (seit 1779,<lb/>
Gohin). — Vosges <i>recomplettei</i> „rendre complet<lb/>
ce qui ne l'était pas“ P.</p>
<p>Nfr. <i>incomplet</i> „non complet“ (seit Ac 1762),<lb/>
npr. <i>incoumplèt</i>. — Mfr. <i>incomplètement</i> „d'une<lb/>
manière incomplète“ (1503, RF 32, 83), nfr. id.<lb/>
(seit Besch 1845), npr. <i>incoumpletomen</i>.</p>
<lb/>
<p>b. Grundwort und ablt. in speziellen berufs-<lb/>
sprachlichen bed.</p>
<p><i>α.</i> Nfr. <i>complet</i> adj. „(t. d'administr. milit.)<lb/>
(corps de troupes) qui atteint le nombre fixé<lb/>
de son effectif“ (seit Montaigne); übertragen<lb/>
nfr. „entièrement occupé (wagon, autobus)“ (seit<lb/>
Lar 1869). Substantiviert nfr. „état (d'un corps<lb/>
de troupes) qui atteint le nombre fixé“ (seit SSi-<lb/>
mon, Lar 1929). Daher nfr. <i>être au (grand) com-<lb/>
plet</i> „atteindre le nombre fixé (en parlant d'un<lb/>
corps de troupes)“ (seit Boiste 1829) ; in weiterer<lb/>
bed. nfr. „id. (en parlant d'une réunion de per-<lb/>
sonnes quelconques, d'une salle, d'un omnibus, etc.)“<lb/>
(seit Littré 1863), npr. <i>es au coumplèt</i> „c'est au<lb/>
complet“. — Nfr. <i>non-complet</i> „vide existant dans<lb/>
l'effectif d'un corps de troupes“ (Ac 1740—Lar<lb/>
1874). Nfr. <i>incomplet</i> (Voltaire 1742—DG, Trév<lb/>
1771). Nfr. <i>décomplet</i> (1793, Brunot 9, 925). —<lb/>
Nfr. <i>compléter</i> „combler les lacunes (d'un corps<lb/>
de troupes)“ (seit Trév 1771). — Nfr. <i>complète-<lb/>
ment</i> „action de compléter (un corps de troupes)“<lb/>
(seit Trév 1752). Brunot 9, 925.</p>
<p><i>β.</i> Nfr. <i>complet</i> „(t. de libr.) (livre) auquel il<lb/>

ne manque pas de feuilles, (ouvrage) auquel il ne<lb/>
manque pas de volumes“ (seit Widerh 1675). Nfr.<lb/>
<i>incomplet</i> „(livre, ouvrage) qui n'est pas complet“<lb/>
(seit Ac 1762 ; substantiviert seit Besch 1845). —<lb/>
Nfr. <i>compléter</i> „ajouter ce qui manque (à un ou-<lb/>
vrage, à une édition)“ (seit 1733, Trév 1771); <i>se<lb/>
compléter</i> „prendre les livraisons destinées à com-<lb/>
pléter un ouvrage“ (Besch 1845—Lar 1869).</p>

<col pg="983" s="a"/>

<p><i>γ.</i> Nfr. <i>complet</i> „(habit) auquel il ne manque<lb/>
aucune des pièces nécessaires“ (seit mitte 17. jh.).<lb/>
Substantiviert „(t. de tailleur) vêtement d'homme<lb/>
dont les diverses pièces sont de la même étoffe“<lb/>
(seit mitte 19. jh.), St-Laur B. <i>k &few-u-boucle; m p l &few-e-ferme-accent;</i> „vête-<lb/>
ment(s)“ ALLo 944. Ironisch Mâcon <i>complet de<lb/>
bois</i> „cercueil“.</p>
<p><i>δ.</i> Nfr. <i>fleur complète</i> „(t. de bot.) fleur qui a<lb/>
un calice, une corolle, une ou plusieurs étamines<lb/>
et un ou plusieurs pistils“ (seit Trév 1771); <i>nymphe<lb/>
incomplète</i> „(t. d'entom.) nymphe qui est pourvue<lb/>
d'ailes et de pattes, mais immobile“ (Besch 1845<lb/>
—Littré 1866). — Nfr. <i>fleur incomplète</i> „f. dé-<lb/>
pourvue de quelque organe, notamment de corolle“<lb/>
(seit Trév 1771). — Nfr. <i>incomplétées</i> „groupe de<lb/>
plantes dont les fleurs sont dépourvues de corolle<lb/>
ou de pétales“ Lar 1931.</p>
<p>2. Vollkommen. — Mfr. nfr. <i>complet</i> „(homme)<lb/>
qui a toutes les qualités désirables ; (joie, succès,<lb/>
etc.) qui ne laisse rien à désirer “ (seit 16. jh.,<lb/>
doch bis 19. jh. selten, Sand), mars. <i>coumplet</i> „par-<lb/>
fait, à qui il ne manque rien“ A. Ironisch in aus-<lb/>
drücken wie nfr. <i>fou complet</i> „personne folle au<lb/>
dernier degré“ (seit Rich 1680) ; mars. <i>es coum-<lb/>
plètto</i> „(t. de mépris) elle est accomplie“ A.</p>
<p>Ablt. — Nfr. <i>compléter (un homme)</i> „le rendre<lb/>
parfait“ (seit Hugo 1835), mars. <i>coumpletar</i> „per-<lb/>
fectionner“ A. — Nfr. <i>décompléter</i> qn „priver qn<lb/>
d'une chose essentielle, dégrader qn“ (Besch 1845<lb/>
—Lar 1870). — Nfr. <i>être incomplet</i> „(t. de rail-<lb/>
lerie) homme incommodé de quelque infirmité“<lb/>
BL 1808, <i>incomplet</i> „imparfait (homme, esprit,<lb/>
joie)“ (Besch 1845—Lar 1873) <e>2</e>). — Nfr. <i>incom-<lb/>
plétude</i> „(t. de méd.) sentiment d'insuffisance, de<lb/>
non-conformité dans les pensées, les émotions et<lb/>
les actions“ Lar 1931.</p>
<lb/>
<p>II. Apr. <i>completiu</i> „qui sert à compléter“ (14.<lb/>

jh.), mfr. <i>complétif</i> (1503, RF 32, 34; 1551, DG,<lb/>
theol. ausdruck; seit 1875, in allg. bed.), npr.<lb/>
<i>completiéu</i>.</p>
<p>III. 1. Afr. <i>complie</i> „dernière partie de l'office<lb/>
qui se dit ou se chante après vêpres“ (12.—14.<lb/>
jh., Gdf; TL; Littré; Brendan W 232; Mon Guill;<lb/>
Pass), fr. <i>complies</i> (seit Chrestien), apr. <i>comple-<lb/>
ta(s)</i> (13. jh.), dauph. <i>coumpleios</i> M, mars. <i>coum-<lb/>
plios</i> A, lang. <i>coumplètos</i> M, aveyr. <i>coumplíos</i>,<lb/>
St-Pierre <i>coumpletas</i>, bearn. <i>coumpletes</i>. — Re-<lb/>
densarten. Mfr. <i>aller à complie</i> „faire l'amour“<lb/>
Ev Quen. Mfr. <i>complie est dicte</i> „c'est fini, tu n'as<lb/>
plus rien à ergoter“ Trepp <e>3</e>).</p>
<p>2. Mfr. <i>completoires</i> „complies“ (16. jh.). —<lb/>
Adjektiviert mfr. <i>completoire</i> „qui appartient aux<lb/>
complies“ (16. jh.).</p>
<lb/>
<p>I ist in beiden bed. entlehnt aus lt. <sc>complētus</sc>,<lb/>
vgl. auch rum. <i>complet</i>, it. sp. pg. <i>completo</i>, kat.<lb/>
<i>complet</i>. Über das verhältnis zu den ablt. von<lb/>
complere s. dieses. Aus dem fr. entlehnt e. <i>com-<lb/>
plete</i>, d. <i>komplett</i> in den meisten bed. des fr.,<lb/>
<col pg="983" s="b"/><lb />
Schulz B, oberhess. <i>cumplet</i> „in etwas vollkommen<lb/>
bescheid wissend“. — II ist entlehnt aus lt. <sc>com-<lb/>
plētivus</sc>, das sich als theologischer ausdruck bei<lb/>
den kirchenschriftstellern, als linguistischer bei<lb/>
Priscian findet. — III 1 setzt kirchenlt. (<sc>hora</sc>)<lb/>
<sc>completa</sc> „letzte gebetsstunde“ fort, mit lautlicher<lb/>
einwirkung von afr. <i>complir</i>, Cohn 225, ebenso<lb/>
piem. <i>cunpía</i>; it. <i>compieta</i>. Im mfr. nfr. gilt nur<lb/>
die pl. form, analog zu <i>matines</i>, <i>laudes</i>, <i>vêpres</i>;<lb/>
ebenso kat. <i>completes</i>, sp. pg. <i>completas</i>. Aus dem<lb/>
fr. entlehnt sind me. <i>cumplie</i>, mndl. <i>complie</i>. III<lb/>
2 ist entlehnt aus kirchenlt. <sc>completorium</sc>. —<lb/>
Hering.</p>
<notes><p>1) Von Trév 1732—1743 noch als ungebräuch-<lb/>
lich bezeichnet.</p>
<p>2) Mfr. <i>incomplet</i> „imparfait“ (1372) in einem<lb/>
aus dem lt. übersetzten text ist aus lt. <sc>incompletus</sc> entlehnt.</p>
<p>3) Ähnliche redensarten, die davon ausgehen,<lb/>
dass die komplet am ende des tages steht, kennt<lb/>
auch das it., Rheinf 369.</p></notes>

<provenance op="Pascale Renders" type="manual" timestamp="1240416000000"/>
</art>
</few>

Section 7 - Compliance

An XML Schema (ffml.xsd) defines the schema of FFML. An XML document can thus be parsed and validated against this schema. Its XML namespace is http://www.atilf.fr/few/ffml

Reference implementations of a validator and a SAX parser handler have been completed (ffml-validator) and are provided by the ATILF research centre. The validator can be used to certify the compliance of an XML document with the definition of FFML. The validator is platform-independent: It can run on e.g. Linux, Mac OS X, MS Windows. It is controlled through a very simple command line interface (e.g. ./ffml-validator name-of-file-to-validate.xml); adding a desktop GUI or web-based interface, although probably unnecessary, is easily doable.

An alternative version of the validator (ffml-schema-validator) is also offered. It checks only the compliance against the FFML Schema, whereas the reference implementation (ffml-validator) also performs supplementary checks against the linguistics rules of what is expected in an FEW article. It is highly recommended to use the reference implementation (ffml-validator).

The remainder of this section can be safely skipped by readers who only wish to use the reference implementation of the validator, but is a must-read for readers who wish to design their own validator.

All stated schema-level constraints but two are enforced by the XML Schema of FFML. If the reference implementation of the validator is used, there is no impact in practice. However, if one wishes to implement her own validator, the following remarks should be considered. The at-least-once-<b>-per-<h> and the at-most-one-<e>-per-<sc> constraints (see Section 3.1 - Font-style Tags) must be enforced by the SAX parser handler, as they are not included in the schema. The reference implementation relies either on Sun's JVM default parser or on Apache Software Foundation's Xerces parser. A benefit of the latter (i.e. Xerces) is the support of RelaxNG Schemas, which could provide a way to enforce the two aforementioned constraints at the schema-level rather than at the handler-level. Indeed, an alternative to obtain full validation at the schema-level may be to define FFML based on a RelaxNG Schema instead of an XML Schema.

Independently from the two aforementioned constraints on font-style tags, the enforcement of the definition of an identifier attribute for <art> tags (see Section 2.1 Dictionary-level Structural Tags) is made at the handler-level, because this attribute is optional by design.

Appendix

This document should always be considered along with its three appendix files:

- ffml.xsd - the formal definition (XML Schema) of FFML
- few-char-map.pdf - character map of FFML files
- completus.pdf - scanned version of the FEW article *completus*

An optional DTD (few-char-map-entity.dtd) of the XML entities derived from the character map can be obtained with the provided tool (few-char-map-entity-dump).

An FFML file can be validated against the defined FFML Schema using the provided tool (ffml-validator). To deactivate supplementary checks against linguistics rules of what is expected in an FEW article, one can also use an alternative version of the tool (ffml-schema-validator).

Multiple articles stored in a given FFML file can be automatically split into separate FFML files using the provided tool (split-ffml-volume).

D.3 Version FFML de l'article CHOCOLATL (FEW 20, 63b)

```
<?xml version="1.0" encoding="UTF-8"?>
<few xmlns="http://www.atilf.fr/few/ffml" volume="20" book="1">
<art id="0" ici="2">
<col pg="63" s="b"/>

<h><b>chocolatl</b> (aztek.) getränk aus kakao.</h>

<p>1. Mfr. <i>chocholate</i> m. „breuvage fait avec des<lb/>
amandes de cacao‟ (1598), <i>chocolate</i> (1640--Trév<lb/>
1732), f. (1640--1672); <i>chocolat</i> m. (seit 1666, Arveil-<lb/>
ler 178); <i>chocolate</i> „substance solide de ce breuvage‟<lb/>
(Rich 1680-Trév 1732), <i>chocolat</i> (seit SavBr 1723),<lb/>
lütt. <i>choûcol&#x00E5;</i> „chocolat‟, loch. <i>choucolâ</i>, sav.<lb/>
<i>chôcolâ</i>, mars. <i>chocoulat</i> A, bearn. <i>chocolat</i>. Saint.<lb/>
SeudreS. <i>chacolat</i>. -- Übertragen Paris <i>chocolat</i><lb/>
„personne, animal, objet de couleur chocolat‟ B;<lb/>
„battu au jeu; n'ayant pas réussi‟ B <e>1</e>).</p>

<p> Ablt. -- Nfr. <i>chocolatière</i> f. „vase où l'on prépare,<lb/>
où l'on sert le chocolat en boisson‟ (seit 1680);<lb/>
<i>chocolatier</i> m. „fabricant, marchand de chocolat‟<lb/>
(seit Rich 1680); <i>chocolaté</i> adj. „qui contient du<lb/>
chocolat (en parlant du café)‟ (seit Trév 1771, 2,<lb/>
148 b).</p>

<p>2. Nfr. <i>chicolate</i> f. „chocolat‟ (1658), Nice <i>chicou-<lb/>
lata</i>; rouchi <i>chicolat</i>, Tourc. art. havr. <i>chucolat</i>, havr.<lb/>
<i>chicolat</i>, Louv. <i>chicoulas</i>, mars. <i>chicoulat</i> A, Vinz.<lb/>
<i>š &#x0259; k &few-o-ferme-long; l &#x00E1;</i>, bearn. <i>chicoulat</i>.</p>
<p>1 ist aus sp. <i>chocolate</i> entlehnt, das selber aus<lb/>
einem aztek. <i>chocolatl</i> übernommen ist. Doch ist<lb/>
dieses selber erst im 19. jh. ausdrücklich belegt, und<lb/>
seine bildung innerhalb des aztek. ist nicht klar, s.<lb/>
dazu Nykl MPh 23, 325; KZ 61, 93; König 66;<lb/>
Friederici 182; Corom 2, 75. Deutlich auf aztek.<lb/>
ursprung weist der erste spanische beleg, von 1580,<lb/>
der <i>chocollatl</i> geschrieben ist. Aus dem sp. auch it.<lb/>
<i>cioccolata</i>, d. <i>schokolade</i>. Wid 1669 berichtet als<lb/>
erster, dass das getränk auch in Spanien und Italien<lb/>
genossen werde, und SavBr 1679, 2, 176 schreibt:<lb/>
<i>le chocolate dont les Francais usent à présent.</i> --<lb/>
Unklar ist auch das verhältnis von 2 zu 1. Der erste<lb/>
beleg von 2 kommt von den kleinen Antillen, wo-<lb/>
durch wahrscheinlich gemacht wird, dass diese form<lb/>
sich selbständig verbreitet hat. Sie lebt auch in<lb/>
piem. <i>cicolata</i>, Poschiavo <i>t š i k &few-u-boucle; l a t a. </i> Merkwürdi-<lb/>
gerweise wird sie im 19. jh. auch für das span. auf<lb/>
```

```
<col pg="64" s="a"/>
den Philippinen bezeugt, unter der form <i>siculate</i>,<lb/>
s. Friederici.</p>
<notes><p>1) Diese bed. ist eine synonymalablt. zu <i>knock-<lb/>
out</i>, das vom volk <i>moka</i> gesprochen wurde. S. noch<lb/>
DauzatArgot 80.</p></notes>

<provenance op="Isabelle Clement" type="manual" timestamp="1275660000000"/>
</art>
</few>
```

Annexe E

FSML

E.1 Introduction

Cette annexe présente l'article CHOCOLATL (FEW 20, 63b) rétroconverti selon le format XML des articles rétroconvertis, dénommé format FSML (pour *FEW semantic-style markup language*). L'analyse du balisage inséré dans cet article est proposée au chapitre 6

Le schéma XML complet correspondant à ce format se trouve sur le CD-ROM qui accompagne cette thèse (→ G.2).

E.2 Version FSML de l'article CHOCOLATL (FEW 20, 63b)

<?xml version="1.0" encoding="UTF-8"?>

<few xmlns="http://www.atilf.fr/few/fsml" book="1" volume="20">

<!-- article map

1 Mfr. chocholate m. „breuvage fait avec des amandes de cacao“ (1598)
1 Ablt. — Nfr. chocolatière f. „vase où l'on prépare, où l'on sert le chocolat en boisson“ (seit 1680)
2 Nfr. chicolate f. „chocolat“ (1658)

-->

<art book="1" ici="2" id="0" lang="german" type="doc-com" volume="20">

<col merge-split-word="no" pg="63" s="b"/>

<entry><etymon lang="aztek." type="vedette">chocolatl</etymon> <lang_etymon>(aztek.)</lang_etymon> getränk
aus kakao.</entry>

<doc>
<p structid="1"><pnum id="1">1.</pnum> <unit><geoling status="ok">Mfr.</geoling> <form><i>chocholate</i></form>
<gram>m.</gram> <def>„breuvage fait avec des<lb merge-split-word="no"/>
amandes de cacao“</def> <precisions status="ok-wraps-parentheses">(<attestation><date>1598</date></attestation>)</
precisions></unit>, <unit><imp contents="Mfr." type="geoling"/><form><i>chocolate</i></form> <imp contents="m."
type="gram"/><imp contents="„breuvage fait avec des amandes de cacao“" type="def"/><precisions status="ok-wraps-
parentheses">(<attestation><date>1640</date>—<biblio status="ok">Trév<lb merge-split-word="no"/>
1732</biblio></attestation>)</precisions></unit>, <unit><imp contents="Mfr." type="geoling"/><imp contents="chocolate"
type="form"/><gram>f.</gram> <imp contents="„breuvage fait avec des amandes de cacao“" type="def"/><precisions
status="ok-wraps-parentheses">(<attestation><date>1640</date>—<date>1672</date></attestation>)</precisions></unit>;
<unit><imp contents="Mfr." type="geoling"/><form><i>chocolat</i></form> <gram>m.</gram> <imp contents="„breuvage
fait avec des amandes de cacao“" type="def"/><precisions status="ok-wraps-parentheses">(<attestation>seit <date>1666</
date>, Arveil-<lb merge-split-word="discard-hyphen"/>
ler 178</attestation>)</precisions></unit>; <unit><imp contents="Mfr." type="geoling"/><form><i>chocolate</i></form>
<imp contents="m." type="gram"/><def>„substance solide de ce breuvage“</def><lb merge-split-word="no"/>
<precisions status="ok-wraps-parentheses">(<attestation><biblio status="ok">Rich 1680</biblio>–<biblio status="ok">Trév
1732</biblio></attestation>)</precisions></unit>, <unit><imp contents="Mfr." type="geoling"/><form><i>chocolat</i></
form> <imp contents="m." type="gram"/><imp contents="„substance solide de ce breuvage“" type="def"/><precisions
status="ok-wraps-parentheses">(<attestation>seit <biblio status="ok">SavBr 1723</biblio></attestation>)</precisions></
unit>,<lb merge-split-word="no"/>
<unit><geoling status="ok">lütt.</geoling> <form><i>choûcolå</i></form> <imp contents="m." type="gram"/
><def>„chocolat“</def></unit>, <unit><geoling status="ok">loch.</geoling> <form><i>choucolâ</i></form></unit>,
<unit><geoling status="ok">sav.</geoling><lb merge-split-word="no"/>
<form><i>chôcolâ</i></form></unit>, <unit><geoling status="ok">mars.</geoling> <form><i>chocoulat</i></form>
<precisions status="ok-wraps-tag"><biblio status="ok">A</biblio></precisions></unit>, <unit><geoling
status="ok">bearn</geoling>. <form><i>chocolat</i></form></unit>. <unit><geoling status="ok">Saint.</geoling><lb
merge-split-word="no"/>
<geoling status="ok">SeudreS.</geoling> <form><i>chacolat</i></form></unit>. —<group><title> Übertragen </
title><unit><geoling status="ok">Paris</geoling> <form><i>chocolat</i></form><lb merge-split-word="no"/>
<def>„personne, animal, objet de couleur chocolat“</def> <precisions status="ok-wraps-tag"><biblio status="ok">B</
biblio></precisions></unit>;<lb merge-split-word="no"/>
<unit><imp contents="Paris" type="geoling"/><imp contents="chocolat" type="form"/><def>„battu au jeu; n'ayant pas
réussi“</def> <precisions status="ok-wraps-tag"><biblio status="ok">B</biblio></precisions></unit> <appelnote id="1"
status="ok"><e>1)</e></appelnote>.</group></p>
</doc>

<doc>
<p headtitle="Ablt. —" structid="1"><group><title>Ablt. —</title> <unit><geoling status="ok">Nfr.</geoling>
<form><i>chocolatière</i></form> <gram>f.</gram> <def>„vase où l'on prépare,<lb merge-split-word="no"/>
où l'on sert le chocolat en boisson“</def> <precisions status="ok-wraps-parentheses">(<attestation>seit <date>1680</date></
attestation>)</precisions></unit>;<lb merge-split-word="no"/>
<unit><imp contents="Nfr." type="geoling"/><form><i>chocolatier</i></form> <gram>m.</gram> <def>„fabricant,
marchand de chocolat“</def><lb merge-split-word="no"/>

<precisions status="ok-wraps-parentheses">(<attestation>seit <biblio status="ok">Rich 1680</biblio></attestation>)</
precisions></unit>; <unit><imp contents="Nfr." type="geoling"/><form><i>chocolaté</i></form> <gram>adj.</gram>
<def>„qui contient du<lb merge-split-word="no"/>
chocolat (en parlant du café)“</def> <precisions status="ok-wraps-parentheses">(<attestation>seit <biblio status="ok">Trév
1771, 2,<lb merge-split-word="no"/>
148</biblio> b</attestation>)</precisions></unit>.</group></p>
</doc>

<doc>
<p structid="2"><pnum id="2">2.</pnum> <unit><geoling status="ok">Nfr.</geoling> <form><i>chicolate</i></form>
<gram>f.</gram> <def>„chocolat“</def> <precisions status="ok-wraps-parentheses">(<attestation><date>1658</date></
attestation>)</precisions></unit>, <unit><geoling status="ok">Nice</geoling> <form><i>chicou-<lb merge-split-
word="no"/>
lata</i></form></unit>; <unit><geoling status="ok">rouchi</geoling> <form><i>chicolat</i></form></unit>,
<unit><geoling status="ok">Tourc.</geoling> <geoling status="ok">art.</geoling> <geoling status="ok">havr.</geoling>
<form><i>chucolat</i></form></unit>, <unit><geoling status="ok">havr.</geoling><lb merge-split-word="no"/>
<form><i>chicolat</i></form></unit>, <unit><geoling status="ok">Louv.</geoling> <form><i>chicoulas</i></form></
unit>, <unit><geoling status="ok">mars.</geoling> <form><i>chicoulat</i></form> <precisions status="ok-wraps-
tag"><biblio status="ok">A</biblio></precisions></unit>, <unit><geoling status="ok">Vinz</geoling>.<lb merge-split-
word="no"/>
<form><i>š ə k [illegible] l á</i></form></unit>, <unit><geoling status="ok">bearn.</geoling> <form><i>chicoulat</i></form></
unit>.</p>
</doc>

<com>
<p><pref id="1" status="ok">1</pref> ist aus <lang status="ok">sp.</lang> <form><i>chocolate</i></form> entlehnt, das
selber aus<lb merge-split-word="no"/>
einem aztek. <i><etymon descendance="emprunts" type="vedette">chocolatl</etymon></i> übernommen ist. <biblio
status="ok">Doch</biblio> ist<lb merge-split-word="no"/>
dieses selber erst im <date>19. jh.</date> ausdrücklich belegt, und<lb merge-split-word="no"/>
seine bildung innerhalb des aztek. ist nicht klar, <gram>s.</gram><lb merge-split-word="no"/>
dazu Nykl <biblio status="ok">MPh 23, 325</biblio>; <biblio status="ok">KZ 61, 93</biblio>; <biblio status="ok">König
66</biblio>;<lb merge-split-word="no"/>
<biblio status="ok">Friederici 182</biblio>; <biblio status="ok">Corom 2, 75</biblio>. Deutlich auf aztek.<lb merge-split-
word="no"/>
ursprung weist der erste spanische beleg, von <date>1580</date>,<lb merge-split-word="no"/>
der <form><i>chocollatl</i></form> geschrieben ist. Aus dem <lang status="ok">sp.</lang> auch <lang status="ok">it.</
lang><lb merge-split-word="no"/>
<form><i>cioccolata</i></form>, <lang status="ok">d.</lang> <form><i>schokolade</i></form>. <biblio
status="geoling?">Wid 1669</biblio> berichtet als<lb merge-split-word="no"/>
erster, dass das getränk auch in Spanien und Italien<lb merge-split-word="no"/>
genossen werde, und <biblio status="ok">SavBr 1679, 2, 176</biblio> schreibt:<lb merge-split-word="no"/>
<form type="citation?"><i>le chocolate dont les Francais usent à présent.</i></form> —<lb merge-split-word="no"/>
Unklar ist auch das verhältnis von <pref id="2" status="ok">2</pref> zu <pref id="1" status="ok">1</pref>. <biblio
status="ok">Der</biblio> erste<lb merge-split-word="no"/>
beleg von <pref id="2" status="ok">2</pref> kommt von den kleinen Antillen, wo-<lb merge-split-word="no"/>
durch wahrscheinlich gemacht wird, dass diese form<lb merge-split-word="no"/>
sich selbständig verbreitet hat. Sie lebt auch in<lb merge-split-word="no"/>
<lang status="ok">piem.</lang> <form><i>cicolata</i></form>, <lang status="ok">Poschiavo</lang> <form><i>t š i k [illegible] l
a t a.</i></form> Merkwürdi-<lb merge-split-word="discard-hyphen"/>
gerweise wird sie im <date>19. jh.</date> auch für das span. auf

<col merge-split-word="no" pg="64" s="a"/>

den Philippinen bezeugt, unter der form <form><i>siculate</i></form>,<lb merge-split-word="no"/>
s. <biblio status="ok">Friederici</biblio>.<signature author="Wartburg" lang="german"/></p>
</com>

<notes call-sequence="ok" count="1">

<p><note id="1">1) Diese bed. ist eine synonymalablt. zu <i><etymon type="?">knock-<lb merge-split-word="keep-
hyphen"/>

out</etymon></i>, das vom volk <form><i>moka</i></form> gesprochen wurde. <biblio status="ok">S</biblio>. noch<lb
merge-split-word="no"/>
<biblio status="ok">DauzatArgot 80</biblio>.</note></p>

</notes>

<provenance op="Isabelle Clement" timestamp="1275660000000" type="manual"/>
<provenance op="identity" timestamp="1297696375817" type="algo"/>
<provenance op="detect-corrupted-entities" timestamp="1297696375821" type="algo"/>
<provenance op="streamline-p-extreme-spaces" timestamp="1297696375830" type="algo">
<provenance_update count="2" tag=" " type="removed text chunks"/>
<provenance_update count="1" tag=" " type="updated text chunks"/>
</provenance>
<provenance op="streamline-void-tags" timestamp="1297696375854" type="algo"/>
<provenance op="streamline-breaks" timestamp="1297696375947" type="algo">
<provenance_update count="1" tag="lb" type="removed tags"/>
<provenance_update count="9" tag=" " type="removed text chunks"/>
</provenance>
<provenance op="merge-split-words" timestamp="1297696375950" type="algo">
<provenance_update count="2" tag="col" type="updated tags"/>
<provenance_update count="37" tag="lb" type="updated tags"/>
</provenance>
<provenance op="streamline-layout-tags" timestamp="1297696375961" type="algo">
<provenance_update count="1" tag=" " type="updated text chunks"/>
<provenance_warning msg="warning: moved trailing space out of <i>t š i k [illegible] l a t a.</i>"/>
</provenance>
<provenance op="streamline-quotes" timestamp="1297696376043" type="algo"/>
<provenance op="detect-dubious-spacing" timestamp="1297696376045" type="algo"/>
<provenance op="tag-notes" timestamp="1297696376078" type="algo">
<provenance_update count="1" tag="notes" type="spliced tags"/>
<provenance_update count="1" tag="note" type="spliced tags"/>
</provenance>
<provenance op="tag-appels-notes" timestamp="1297696376266" type="algo">
<provenance_update count="1" tag="appelnote" type="spliced tags"/>
<provenance_update count="1" tag="notes" type="updated tags"/>
<provenance_update count="1" tag=" " type="updated text chunks"/>
</provenance>
<provenance op="tag-entry" timestamp="1297696376475" type="algo">
<provenance_update count="1" tag="entry" type="updated tags"/>
</provenance>
<provenance op="tag-etymon" timestamp="1297696376766" type="algo">
<provenance_update count="4" tag="etymon" type="spliced tags"/>
</provenance>
<provenance op="tag-lang-etymon" timestamp="1297696377001" type="algo">
<provenance_update count="1" tag="lang_etymon" type="spliced tags"/>
<provenance_update count="1" tag="etymon" type="updated tags"/>
</provenance>
<provenance op="tag-renvoi" timestamp="1297696377140" type="algo"/>
<provenance op="tag-signature" timestamp="1297696377337" type="algo">
<provenance_update count="1" tag="signature" type="spliced tags"/>
<provenance_update count="1" tag="art" type="updated tags"/>
</provenance>
<provenance op="tag-def" timestamp="1297696377462" type="algo">
<provenance_update count="9" tag="def" type="spliced tags"/>
</provenance>
<provenance op="tag-biblio" timestamp="1297696377777" type="algo">
<provenance_update count="22" tag="biblio" type="spliced tags"/>
<provenance_warning msg="warning: cannot decide whether keyword |Wid| should be tagged with <geoling> or
<biblio> "/>
</provenance>
<provenance op="tag-geoling" timestamp="1297696377954" type="algo">
<provenance_update count="21" tag="geoling" type="spliced tags"/>

<provenance_update count="6" tag="lang" type="spliced tags"/>
</provenance>
<provenance op="streamline-void-tags" timestamp="1297696378809" type="algo"/>
<provenance op="split-doc-com" timestamp="1297696378823" type="algo">
<provenance_update count="1" tag="com" type="spliced tags"/>
<provenance_update count="3" tag="doc" type="spliced tags"/>
<provenance_update count="1" tag="art" type="updated tags"/>
</provenance>
<provenance op="tag-numbering" timestamp="1297696378868" type="algo">
<provenance_update count="2" tag="pnum" type="spliced tags"/>
<provenance_update count="4" tag="pref" type="spliced tags"/>
</provenance>
<provenance op="split-mixt-art" timestamp="1297696379021" type="algo"/>
<provenance op="tag-affix" timestamp="1297696379336" type="algo"/>
<provenance op="tag-date" timestamp="1297696379699" type="algo">
<provenance_update count="10" tag="date" type="spliced tags"/>
</provenance>
<provenance op="tag-gram" timestamp="1297696380293" type="algo">
<provenance_update count="9" tag="gram" type="spliced tags"/>
<provenance_update count="1" tag="gram" type="removed tags"/>
<provenance_warning msg="warning: disambiguated <gram>s.</gram> before <biblio>"/>
</provenance>
<provenance op="tag-form" timestamp="1297696380647" type="algo">
<provenance_update count="33" tag="form" type="spliced tags"/>
<provenance_update count="1" tag="etymon" type="removed tags"/>
</provenance>
<provenance op="tag-concept" timestamp="1297696380932" type="algo"/>
<provenance op="tag-precisions" timestamp="1297696381375" type="algo">
<provenance_update count="14" tag="precisions" type="spliced tags"/>
</provenance>
<provenance op="tag-attestation" timestamp="1297696381433" type="algo">
<provenance_update count="10" tag="attestation" type="spliced tags"/>
</provenance>
<provenance op="tag-title" timestamp="1297696382043" type="algo">
<provenance_update count="2" tag="title" type="spliced tags"/>
</provenance>
<provenance op="tag-unit" timestamp="1297696382445" type="algo">
<provenance_update count="18" tag="imp" type="spliced tags"/>
<provenance_update count="26" tag="unit" type="spliced tags"/>
<provenance_warning msg="warning: no <precisions> found in unit | <geoling>lütt.</geoling> <form>choûcolâ</form> <imp type='gram' contents='m.' /> <def>„chocolat“</def>|"/>
<provenance_warning msg="warning: no <precisions> found in unit | <geoling>loch.</geoling> <form>choucolâ</form>|"/>
<provenance_warning msg="warning: no <precisions> found in unit | <geoling>sav.</geoling> <form>chôcolâ</form>|"/>
<provenance_warning msg="warning: no <precisions> found in unit | <geoling>bearn</geoling> <form>chocolat</form>|"/>
<provenance_warning msg="warning: no <precisions> found in unit | <geoling>Saint.</geoling> <geoling>SeudreS.</geoling> <form>chacolat</form>|"/>
<provenance_warning msg="warning: no <precisions> found in unit | <geoling>Nice</geoling> <form>chicou- lata</form>|"/>
<provenance_warning msg="warning: no <precisions> found in unit | <geoling>rouchi</geoling> <form>chicolat</form>|"/>
<provenance_warning msg="warning: no <precisions> found in unit | <geoling>Tourc.</geoling> <geoling>art.</geoling> <geoling>havr.</geoling> <form>chucolat</form>|"/>
<provenance_warning msg="warning: no <precisions> found in unit | <geoling>havr.</geoling> <form>chicolat</form>|"/>
<provenance_warning msg="warning: no <precisions> found in unit | <geoling>Louv.</geoling> <form>chicoulas</form>|"/>
<provenance_warning msg="warning: no <precisions> found in unit | <geoling>Vinz</geoling> <form>š ə k [illegible] l á</form>|"/>
<provenance_warning msg="warning: no <precisions> found in unit | <geoling>bearn.</geoling>

<form>chicoulat</form>|"/>
</provenance>
<provenance op="merge-mixt-p" timestamp="1297696382697" type="algo"/>
<provenance op="tag-microstructure" timestamp="1297696382704" type="algo">
<provenance_update count="2" tag="group" type="spliced tags"/>
<provenance_update count="3" tag="p" type="updated tags"/>
<provenance_update count="1" tag=" " type="spliced text chunks"/>
</provenance>
<provenance op="analyze-chunk-stream" timestamp="1297696382947" type="algo"/>
<provenance op="export-chunk-stream-trace" timestamp="1297696383086" type="algo"/>
<provenance op="show-tags" timestamp="1297696384354" type="algo"/>
<provenance op="show-isolated-tags" timestamp="1297696384460" type="algo"/>
<provenance op="show-untagged-text" timestamp="1297696384472" type="algo"/>
<provenance op="show-untagged-unit-text" timestamp="1297696384660" type="algo"/>
<provenance op="show-frequencies" timestamp="1297696384675" type="algo">
<provenance_update count="1" tag="art" type="FFML tag frequency"/>
<provenance_update count="1" tag="b" type="FFML tag frequency"/>
<provenance_update count="2" tag="col" type="FFML tag frequency"/>
<provenance_update count="1" tag="few" type="FFML tag frequency"/>
<provenance_update count="1" tag="e" type="FFML tag frequency"/>
<provenance_update count="0" tag="h" type="FFML tag frequency"/>
<provenance_update count="35" tag="i" type="FFML tag frequency"/>
<provenance_update count="37" tag="lb" type="FFML tag frequency"/>
<provenance_update count="1" tag="notes" type="FFML tag frequency"/>
<provenance_update count="5" tag="p" type="FFML tag frequency"/>
<provenance_update count="41" tag="provenance" type="FFML tag frequency"/>
<provenance_update count="0" tag="sc" type="FFML tag frequency"/>
<provenance_update count="0" tag="provenance_exception" type="FSML tag frequency"/>
<provenance_update count="40" tag="provenance_update" type="FSML tag frequency"/>
<provenance_update count="15" tag="provenance_warning" type="FSML tag frequency"/>
<provenance_update count="0" tag="affix" type="FSML tag frequency"/>
<provenance_update count="1" tag="appelnote" type="FSML tag frequency"/>
<provenance_update count="10" tag="attestation" type="FSML tag frequency"/>
<provenance_update count="22" tag="biblio" type="FSML tag frequency"/>
<provenance_update count="1" tag="com" type="FSML tag frequency"/>
<provenance_update count="0" tag="concept" type="FSML tag frequency"/>
<provenance_update count="10" tag="date" type="FSML tag frequency"/>
<provenance_update count="9" tag="def" type="FSML tag frequency"/>
<provenance_update count="3" tag="doc" type="FSML tag frequency"/>
<provenance_update count="1" tag="entry" type="FSML tag frequency"/>
<provenance_update count="3" tag="etymon" type="FSML tag frequency"/>
<provenance_update count="33" tag="form" type="FSML tag frequency"/>
<provenance_update count="21" tag="geoling" type="FSML tag frequency"/>
<provenance_update count="8" tag="gram" type="FSML tag frequency"/>
<provenance_update count="18" tag="imp" type="FSML tag frequency"/>
<provenance_update count="0" tag="k" type="FSML tag frequency"/>
<provenance_update count="6" tag="lang" type="FSML tag frequency"/>
<provenance_update count="1" tag="lang_etymon" type="FSML tag frequency"/>
<provenance_update count="0" tag="milestone" type="FSML tag frequency"/>
<provenance_update count="0" tag="mixt" type="FSML tag frequency"/>
<provenance_update count="1" tag="note" type="FSML tag frequency"/>
<provenance_update count="0" tag="part_com" type="FSML tag frequency"/>
<provenance_update count="2" tag="pnum" type="FSML tag frequency"/>
<provenance_update count="14" tag="precisions" type="FSML tag frequency"/>
<provenance_update count="4" tag="pref" type="FSML tag frequency"/>
<provenance_update count="0" tag="renvoi" type="FSML tag frequency"/>
<provenance_update count="1" tag="signature" type="FSML tag frequency"/>
<provenance_update count="2" tag="group" type="FSML tag frequency"/>
<provenance_update count="2" tag="title" type="FSML tag frequency"/>
<provenance_update count="26" tag="unit" type="FSML tag frequency"/>
</provenance>

</art>
<!-- show-frequencies

337 words

2072 characters

379 XML elements

125 FFML elements

1 x <art>
1 x <b>
2 x <col>
1 x <few>
1 x <e>
0 x <h>
35 x <i>
37 x <lb>
1 x <notes>
5 x <p>
41 x <provenance>
0 x <sc>

254 FSML elements

0 x <provenance_exception>
40 x <provenance_update>
15 x <provenance_warning>
0 x <affix>
1 x <appelnote>
10 x <attestation>
22 x <biblio>
1 x <com>
0 x <concept>
10 x <date>
9 x <def>
3 x <doc>
1 x <entry>
3 x <etymon>
33 x <form>
21 x <geoling>
8 x <gram>
18 x <imp>
0 x <k>
6 x <lang>
1 x <lang_etymon>
0 x <milestone>
0 x <mixt>
1 x <note>
0 x <part_com>
2 x <pnum>
14 x <precisions>
4 x <pref>
0 x <renvoi>
1 x <signature>
2 x <group>
2 x <title>
26 x <unit>

-->

<!-- show-untagged-unit-text

Untagged text in paragraph units, column 63b

Untagged text in paragraph units, column 63b

Untagged text in paragraph units, column 63b

Untagged text in paragraph units, column 63b

Untagged text in paragraph units, column 64a

Average tagging rate over units of the article: 100%

-->

<!-- show-untagged-text

Untagged text in entry, column 63b (tagging rate: 51%)

<...> getränk aus kakao.

Untagged text in <doc> paragraph, column 63b (tagging rate: 100%)

<...>

Untagged text in <doc> paragraph, column 63b (tagging rate: 100%)

<...>

Untagged text in <doc> paragraph, column 63b (tagging rate: 100%)

Untagged text in <com> paragraph, column 63b (tagging rate: 35%)

<...> ist aus <...> entlehnt, das selber aus<...>einem aztek. <...> übernommen ist. <...> ist<...>dieses selber erst im <...> ausdrücklich belegt, und<...>seine bildung innerhalb des aztek. ist nicht klar, <...>dazu Nykl <...>. Deutlich auf aztek.<...>ursprung weist der erste spanische beleg, von <...>der <...> geschrieben ist. Aus dem <...> auch <...> berichtet als<...>erster, dass das getränk auch in Spanien und Italien<...>genossen werde, und <...> schreibt:<...>Unklar ist auch das verhältnis von <...> zu <...> erste<...>beleg von <...> kommt von den kleinen Antillen, wo-<...>durch wahrscheinlich gemacht wird, dass diese form<...>sich selbständig verbreitet hat. Sie lebt auch in<...> Merkwürdi-<...>gerweise wird sie im <...> auch für das span. auf<...>den Philippinen bezeugt, unter der form <...>s.<...>

Untagged text in <notes> paragraph, column 64a (tagging rate: 31%)

<...>1) Diese bed. ist eine synonymalablt. zu <...>, das vom volk <...> gesprochen wurde. <...>. noch<...>

Average tagging rate over relevant sections of the article: 100%

-->

<!-- show-isolated-tags

Isolated tags found in <doc> paragraph, column 63b (0 isolated tag)

Isolated tags found in <doc> paragraph, column 63b (0 isolated tag)

Isolated tags found in <doc> paragraph, column 63b (0 isolated tag)

-->

<!-- show-tags

Tags inserted into entry, column 63b

chocolatl
<etymon>chocolatl</etymon>
<lang_etymon>(aztek.)</lang_etymon>

Tags inserted into <p> paragraph, column 63b

<pnum>1.</pnum>

<unit>Mfr. chocholate m. „breuvage fait avec des amandes de cacao“ (1598)</unit>
<geoling>Mfr.</geoling>
<form>chocholate</form>
<i>chocholate</i>
<gram>m.</gram>
<def>„breuvage fait avec des amandes de cacao“</def>
<lb/>
<precisions>(1598)</precisions>
<attestation>1598</attestation>
<date>1598</date>

<unit>chocolate (1640—Trév 1732)</unit>
<imp/>
<form>chocolate</form>
<i>chocolate</i>
<imp/>
<imp/>
<precisions>(1640—Trév 1732)</precisions>
<attestation>1640—Trév 1732</attestation>
<date>1640</date>
<biblio>Trév 1732</biblio>
<lb/>

<unit>f. (1640—1672)</unit>
<imp/>
<imp/>
<gram>f.</gram>
<imp/>
<precisions>(1640—1672)</precisions>
<attestation>1640—1672</attestation>
<date>1640</date>
<date>1672</date>

<unit>chocolat m. (seit 1666, Arveiller 178)</unit>
<imp/>
<form>chocolat</form>
<i>chocolat</i>
<gram>m.</gram>
<imp/>
<precisions>(seit 1666, Arveiller 178)</precisions>
<attestation>seit 1666, Arveiller 178</attestation>
<date>1666</date>

<lb/>

<unit>chocolate „substance solide de ce breuvage“ (Rich 1680–Trév 1732)</unit>
<imp/>
<form>chocolate</form>
<i>chocolate</i>
<imp/>
<def>„substance solide de ce breuvage“</def>
<lb/>
<precisions>(Rich 1680–Trév 1732)</precisions>
<attestation>Rich 1680–Trév 1732</attestation>
<biblio>Rich 1680</biblio>
<biblio>Trév 1732</biblio>

<unit>chocolat (seit SavBr 1723)</unit>
<imp/>
<form>chocolat</form>
<i>chocolat</i>
<imp/>
<imp/>
<precisions>(seit SavBr 1723)</precisions>
<attestation>seit SavBr 1723</attestation>
<biblio>SavBr 1723</biblio>
<lb/>

<unit>lütt. choûcolå „chocolat“</unit>
<geoling>lütt.</geoling>
<form>choûcolå</form>
<i>choûcolå</i>
<imp/>
<def>„chocolat“</def>

<unit>loch. choucolâ</unit>
<geoling>loch.</geoling>
<form>choucolâ</form>
<i>choucolâ</i>

<unit>sav. chôcolâ</unit>
<geoling>sav.</geoling>
<lb/>
<form>chôcolâ</form>
<i>chôcolâ</i>

<unit>mars. chocoulat A</unit>
<geoling>mars.</geoling>
<form>chocoulat</form>
<i>chocoulat</i>
<precisions>A</precisions>
<biblio>A</biblio>

<unit>bearn. chocolat</unit>
<geoling>bearn</geoling>
<form>chocolat</form>
<i>chocolat</i>

<unit>Saint. SeudreS. chacolat</unit>
<geoling>Saint.</geoling>
<lb/>
<geoling>SeudreS.</geoling>
<form>chacolat</form>
<i>chacolat</i>
<group> Übertragen Paris chocolat „personne, animal, objet de couleur chocolat“ B; „battu au jeu; n'ayant pas réussi“ B

1).</group>
<title> Übertragen </title>

<unit>Paris chocolat „personne, animal, objet de couleur chocolat“ B</unit>
<geoling>Paris</geoling>
<form>chocolat</form>
<i>chocolat</i>
<lb/>
<def>„personne, animal, objet de couleur chocolat“</def>
<precisions>B</precisions>
<biblio>B</biblio>
<lb/>

<unit>„battu au jeu; n'ayant pas réussi“ B</unit>
<imp/>
<imp/>
<def>„battu au jeu; n'ayant pas réussi“</def>
<precisions>B</precisions>
<biblio>B</biblio>
<appelnote>1)</appelnote>
<e>1)</e>

Tags inserted into <p> paragraph, column 63b

<group>Ablt. — Nfr. chocolatière f. „vase où l'on prépare, où l'on sert le chocolat en boisson“ (seit 1680); chocolatier m. „fabricant, marchand de chocolat“ (seit Rich 1680); chocolaté adj. „qui contient du chocolat (en parlant du café)“ (seit Trév 1771, 2, 148 b).</group>
<title>Ablt. —</title>

<unit>Nfr. chocolatière f. „vase où l'on prépare, où l'on sert le chocolat en boisson“ (seit 1680)</unit>
<geoling>Nfr.</geoling>
<form>chocolatière</form>
<i>chocolatière</i>
<gram>f.</gram>
<def>„vase où l'on prépare, où l'on sert le chocolat en boisson“</def>
<lb/>
<precisions>(seit 1680)</precisions>
<attestation>seit 1680</attestation>
<date>1680</date>
<lb/>

<unit>chocolatier m. „fabricant, marchand de chocolat“ (seit Rich 1680)</unit>
<imp/>
<form>chocolatier</form>
<i>chocolatier</i>
<gram>m.</gram>
<def>„fabricant, marchand de chocolat“</def>
<lb/>
<precisions>(seit Rich 1680)</precisions>
<attestation>seit Rich 1680</attestation>
<biblio>Rich 1680</biblio>

<unit>chocolaté adj. „qui contient du chocolat (en parlant du café)“ (seit Trév 1771, 2, 148 b)</unit>
<imp/>
<form>chocolaté</form>
<i>chocolaté</i>
<gram>adj.</gram>
<def>„qui contient du chocolat (en parlant du café)“</def>
<lb/>
<precisions>(seit Trév 1771, 2, 148 b)</precisions>
<attestation>seit Trév 1771, 2, 148 b</attestation>
<biblio>Trév 1771, 2, 148</biblio>

<lb/>

Tags inserted into <p> paragraph, column 63b

<pnum>2.</pnum>

<unit>Nfr. chicolate f. „chocolat“ (1658)</unit>
<geoling>Nfr.</geoling>
<form>chicolate</form>
<i>chicolate</i>
<gram>f.</gram>
<def>„chocolat“</def>
<precisions>(1658)</precisions>
<attestation>1658</attestation>
<date>1658</date>

<unit>Nice chicou- lata</unit>
<geoling>Nice</geoling>
<form>chicou- lata</form>
<i>chicou- lata</i>
<lb/>

<unit>rouchi chicolat</unit>
<geoling>rouchi</geoling>
<form>chicolat</form>
<i>chicolat</i>

<unit>Tourc. art. havr. chucolat</unit>
<geoling>Tourc.</geoling>
<geoling>art.</geoling>
<geoling>havr.</geoling>
<form>chucolat</form>
<i>chucolat</i>

<unit>havr. chicolat</unit>
<geoling>havr.</geoling>
<lb/>
<form>chicolat</form>
<i>chicolat</i>

<unit>Louv. chicoulas</unit>
<geoling>Louv.</geoling>
<form>chicoulas</form>
<i>chicoulas</i>

<unit>mars. chicoulat A</unit>
<geoling>mars.</geoling>
<form>chicoulat</form>
<i>chicoulat</i>
<precisions>A</precisions>
<biblio>A</biblio>

<unit>Vinz. š ə k [illegible] l á</unit>
<geoling>Vinz</geoling>
<lb/>
<form>š ə k [illegible] l á</form>
<i>š ə k [illegible] l á</i>

<unit>bearn. chicoulat</unit>
<geoling>bearn.</geoling>
<form>chicoulat</form>
<i>chicoulat</i>

Tags inserted into <p> paragraph, column 63b

<pref>1</pref>
<lang>sp.</lang>
<form>chocolate</form>
<i>chocolate</i>
<lb/>
<i>chocolatl</i>
<etymon>chocolatl</etymon>
<biblio>Doch</biblio>
<lb/>
<date>19. jh.</date>
<lb/>
<gram>s.</gram>
<lb/>
<biblio>MPh 23, 325</biblio>
<biblio>KZ 61, 93</biblio>
<biblio>König 66</biblio>
<lb/>
<biblio>Friederici 182</biblio>
<biblio>Corom 2, 75</biblio>
<lb/>
<date>1580</date>
<lb/>
<form>chocollatl</form>
<i>chocollatl</i>
<lang>sp.</lang>
<lang>it.</lang>
<lb/>
<form>cioccolata</form>
<i>cioccolata</i>
<lang>d.</lang>
<form>schokolade</form>
<i>schokolade</i>
<biblio>Wid 1669</biblio>
<lb/>
<lb/>
<biblio>SavBr 1679, 2, 176</biblio>
<lb/>
<form>le chocolate dont les Francais usent à présent.</form>
<i>le chocolate dont les Francais usent à présent.</i>
<lb/>
<pref>2</pref>
<pref>1</pref>
<biblio>Der</biblio>
<lb/>
<pref>2</pref>
<lb/>
<lb/>
<lb/>
<lang>piem.</lang>
<form>cicolata</form>
<i>cicolata</i>
<lang>Poschiavo</lang>
<form>t š i k [illegible] l a t a.</form>
<i>t š i k [illegible] l a t a.</i>
<lb/>
<date>19. jh.</date>
<col/>
<form>siculate</form>
<i>siculate</i>

<lb/>
<biblio>Friederici</biblio>
<signature/>

Tags inserted into <p> paragraph, column 64a

<note>1) Diese bed. ist eine synonymalablt. zu knock-out, das vom volk moka gesprochen wurde. S. noch DauzatArgot 80.</note>
<i>knock-out</i>
<etymon>knock-out</etymon>
<lb/>
<form>moka</form>
<i>moka</i>
<biblio>S</biblio>
<lb/>
<biblio>DauzatArgot 80</biblio>

-->

<!-- export-units-to-xs

<fiche etymon="chocolatl" lang="Mfr." forme="chocholate" gram="m." def="„breuvage fait avec des amandes de cacao“" ref="(1598)" N="FEW 20/1, 63b, ici 2, §1, u1"></fiche>

<fiche etymon="chocolatl" lang="(imp.) Mfr." forme="chocolate" gram="(imp.) m." def="(imp.) „breuvage fait avec des amandes de cacao“" ref="(1640—Trév 1732)" N="FEW 20/1, 63b, ici 2, §1, u2"></fiche>

<fiche etymon="chocolatl" lang="(imp.) Mfr." forme="(imp.) chocolate" gram="f." def="(imp.) „breuvage fait avec des amandes de cacao“" ref="(1640—1672)" N="FEW 20/1, 63b, ici 2, §1, u3"></fiche>

<fiche etymon="chocolatl" lang="(imp.) Mfr." forme="chocolat" gram="m." def="(imp.) „breuvage fait avec des amandes de cacao“" ref="(seit 1666, Arveiller 178)" N="FEW 20/1, 63b, ici 2, §1, u4"></fiche>

<fiche etymon="chocolatl" lang="(imp.) Mfr." forme="chocolate" gram="(imp.) m." def="„substance solide de ce breuvage“" ref="(Rich 1680–Trév 1732)" N="FEW 20/1, 63b, ici 2, §1, u5"></fiche>

<fiche etymon="chocolatl" lang="(imp.) Mfr." forme="chocolat" gram="(imp.) m." def="(imp.) „substance solide de ce breuvage“" ref="(seit SavBr 1723)" N="FEW 20/1, 63b, ici 2, §1, u6"></fiche>

<fiche etymon="chocolatl" lang="lütt." forme="choûcolå" gram="(imp.) m." def="„chocolat“" N="FEW 20/1, 63b, ici 2, §1, u7"></fiche>

<fiche etymon="chocolatl" lang="loch." forme="choucolâ" N="FEW 20/1, 63b, ici 2, §1, u8"></fiche>

<fiche etymon="chocolatl" lang="sav." forme="chôcolâ" N="FEW 20/1, 63b, ici 2, §1, u9"></fiche>

<fiche etymon="chocolatl" lang="mars." forme="chocoulat" ref="A" N="FEW 20/1, 63b, ici 2, §1, u10"></fiche>

<fiche etymon="chocolatl" lang="bearn" forme="chocolat" N="FEW 20/1, 63b, ici 2, §1, u11"></fiche>

<fiche etymon="chocolatl" lang="Saint." lang="SeudreS." forme="chacolat" N="FEW 20/1, 63b, ici 2, §1, u12"></fiche>

<fiche etymon="chocolatl" lang="Paris" forme="chocolat" def="„personne, animal, objet de couleur chocolat“" ref="B" N="FEW 20/1, 63b, ici 2, §1, u13"></fiche>

<fiche etymon="chocolatl" lang="(imp.) Paris" forme="(imp.) chocolat" def="„battu au jeu; n'ayant pas réussi“" ref="B" N="FEW 20/1, 63b, ici 2, §1, u14"></fiche>

<fiche etymon="chocolatl" lang="Nfr." forme="chocolatière" gram="f." def="„vase où l'on prépare, où l'on sert le chocolat en boisson“" ref="(seit 1680)" N="FEW 20/1, 63b, ici 2, §2, u1"></fiche>

<fiche etymon="chocolatl" lang="(imp.) Nfr." forme="chocolatier" gram="m." def="„fabricant, marchand de chocolat“" ref="(seit Rich 1680)" N="FEW 20/1, 63b, ici 2, §2, u2"></fiche>

<fiche etymon="chocolatl" lang="(imp.) Nfr." forme="chocolaté" gram="adj." def="„qui contient du chocolat (en parlant du café)“" ref="(seit Trév 1771, 2, 148 b)" N="FEW 20/1, 63b, ici 2, §2, u3"></fiche>

<fiche etymon="chocolatl" lang="Nfr." forme="chicolate" gram="f." def="„chocolat“" ref="(1658)" N="FEW 20/1, 63b, ici 2, §3, u1"></fiche>

<fiche etymon="chocolatl" lang="Nice" forme="chicou- lata" N="FEW 20/1, 63b, ici 2, §3, u2"></fiche>

<fiche etymon="chocolatl" lang="rouchi" forme="chicolat" N="FEW 20/1, 63b, ici 2, §3, u3"></fiche>

<fiche etymon="chocolatl" lang="Tourc." lang="art." lang="havr." forme="chucolat" N="FEW 20/1, 63b, ici 2, §3, u4"></fiche>

<fiche etymon="chocolatl" lang="havr." forme="chicolat" N="FEW 20/1, 63b, ici 2, §3, u5"></fiche>

<fiche etymon="chocolatl" lang="Louv." forme="chicoulas" N="FEW 20/1, 63b, ici 2, §3, u6"></fiche>

<fiche etymon="chocolatl" lang="mars." forme="chicoulat" ref="A" N="FEW 20/1, 63b, ici 2, §3, u7"></fiche>

<fiche etymon="chocolatl" lang="Vinz" forme="š ə k [illegible] l á" N="FEW 20/1, 63b, ici 2, §3, u8"></fiche>

<fiche etymon="chocolatl" lang="bearn." forme="chicoulat" N="FEW 20/1, 63b, ici 2, §3, u9"></fiche>

-->

</few>

<!--
==
Generated by ffml2fsml on Mon Feb 14 16:13:04 CET 2011
==
FEW Retroconversion Algorithms (few-ra) v0.9 (beta) (rev20110214161025)
(c) 2008-2009 ATILF (CNRS & Nancy-Université)
==
-->

Annexe F

Algorithmes

F.1 Introduction

Cette annexe présente, de façon formelle, la majorité des algorithmes expliqués dans le chapitre 5. Le niveau de détail et d'abstraction dans la description formelle des algorithmes ont été choisis de façon à suffire à une compréhension du raisonnement linguistique mis en œuvre. Tout ce qui n'est pas explicité concerne des détails laissés au choix de l'implémentation. Suivant ce principe, certains algorithmes, considérés comme purement techniques, ne sont pas exposés ci-dessous : il s'agit des algorithmes de prétraitement (à l'exception de *streamline-quotes* et *merge-split-words*), des algorithmes de post-traitement et des algorithmes *merge-mixt-p* et *tag-entry*.

Les algorithmes sont présentés ci-dessous par ordre alphabétique. En introduction se trouvent quelques explications concernant le métalangage utilisé, inspiré du métalangage des commandes gardées de Dijkstra (Dijkstra 1976).

F.1.1 Commandes

Le métalangage des commandes gardées de Dijkstra est basé sur des instructions de répétition et de choix. Ces instructions sont obligatoirement gardées par une expression booléenne (c'est-à-dire une expression dont la valeur vraie ou fausse est vérifiable). Pour que la suite d'instructions soit activée, l'expression booléenne doit être vérifiée.

– DO et OD délimitent une commande de répétition. L'expression booléenne qui en est le gardien est délimitée à gauche par DO, à droite par une flèche (→).

– IF et FI délimitent une commande de choix. Les alternatives sont introduites par []. L'expression booléenne qui en constitue le gardien est délimitée à gauche par IF (ou []), à droite par une flèche (→).

– SKIP, apparaissant à l'intérieur d'une commande de choix, indique qu'aucun traitement n'est appliqué.

– Au sein d'une même séquence d'instructions, chaque instruction est séparée de la précédente par un point-virgule.

– " :=" indique une instruction d'affectation.

– "=", dans une expression booléenne, indique une égalité ; " !=" désigne une non-égalité.

F.1.2 Présentation des instructions

Dans le métalangage utilisé, un algorithme est composé d'un programme principal, délimité par BEGIN et END. Toutes les opérations ne sont pas obligatoirement détaillées au sein de ce programme : plusieurs sont résumées par une proposition placée entre guillemets. Certaines de ces propositions sont reprises et développées à la suite du programme principal : nous parlerons de « sous-programme ». Ces sous-programmes sont délimités par les marques |[et]|.

Les propositions qui ne font pas l'objet d'un développement sont considérées comme « techniques » : leurs détails peuvent être définis librement lors de l'implémentation effective de l'algorithme.

La partition définissant les propriétés des balises lors de la création d'une chaîne virtuelle (→ 4.4.2.3) est présentée sous l'intitulé « partition() ». Elle peut recevoir des arguments.

F.1.3 Typage des variables

Dans le métalangage utilisé, la portée des variables est toujours globale. Toutefois, nous ne déclarons pas toutes les variables en début de programme, mais dans le sous-programme où elles sont utilisées.

Les types de variables utilisés sont les suivants :

1. Types de base : *boolean*, *char*, *integer*, *string* ;
2. Types de base spécifiques au FEW :
   – le type *chunk* représente un chunk (*text chunk* ou *tag chunk* ; concepts définis au chapitre 4.4.2.2) ;
   – le type *tag* représente une balise XML appartenant aux balises autorisées (c'est-à-dire définies dans les schémas FFML et FSML) ;
   – le type *regex* représente une expression régulière (→ 5.2.2.2), conventionnellement exprimée entre deux slashs (/) ;
   – le type *virtualString* représente une chaîne de caractères virtuelle (→ 4.4.2.3 ; 5.2.2.3).
3. Types composés : *set*.

Parmi les variables de type *string*, une variable dont le nom commence par *mr* s'appliquera par convention à une chaîne de caractères retournée par la fonction de recherche d'une expression régulière.

F.2 Merge-split-words

```
|[
VAR
    ck : chunk ;
    bal_licites : set of tags ;
    merge_attr : string ;
    left_text : string ;
    right_text : string
BEGIN
bal_licites := [<lb/>, <col/>]
; "initialiser ck à la première balise de l'article appartenant
   à bal_licites"
DO ck != null →
    left_text := "récupérer le chunk de texte juste à gauche de ck"
    ; right_text := "récupérer le chunk de texte juste à droite de ck"
    ; merge_attr := "définir la valeur de l'attribut merge-split-words"
    ; "ajouter à la balise où se trouve ck un attribut merge-split-word,
        prenant la valeur de la variable merge_attr"
    ; "déplacer ck à la prochaine balise de l'article appartenant
       à bal_licites"
OD
END
]|
```

F.2.1 "définir la valeur de l'attribut merge-split-words"

La commande `"définir la valeur de l'attribut merge-split-words"` nécessite la déclaration de variables supplémentaires :

```
    c : char ;
    left_part : string ;
    right_part : string ;
    liste_tirets : set of char
```

```
|[
liste_tirets := ['-','-','--'] {tiret, tiret cadratin, tiret semi-cadratin}
; IF (left_text = null) OR (right_text = null) →
    {il n'y a pas de chunk de texte avant ou après break_tag}
    merge_attr := "no"
  [] (left_text != null) AND (right_text != null) →
    c := "récupérer le dernier caractère de left_text"
    ; IF c not in liste_tirets →
        merge_attr := "no"
      [] c in liste_tirets →
        {il y a du texte autour de break_tag et il y a un tiret avant
         break_tag}
        left_part := "extraire le dernier mot de left_text"
```

```
        {= mot situé juste avant le tiret}
        ; IF left_part = null →
            {il y a un caractère délimiteur juste avant le tiret}
            merge_attr := "no"
          [] left_part != null →
            right_part := "extraire le premier mot de right_text"
            {= mot situé juste après ck}
            ; "comparer avec les mots-clés et définir merge-split-words"
        FI
    FI
FI
]|
```

F.2.2 "comparer avec les mots-clés et définir merge-split-words"

La commande `"définir la valeur de l'attribut merge-split-words"` nécessite la déclaration de variables supplémentaires :

```
    merged_word_with_dash : string ;
    merged_word_without_dash : string ;
    left_part_without_dash : string ;
    mr, mr_left, mr_right : string ;
    keyword_with_dash : boolean ;
    keyword_without_dash : boolean ;
    is_part_a_keyword : boolean

|[
{fusionner les deux mots avec le tiret et chercher si ça forme un mot-clé}
merged_word_with_dash := left_part + right_part
; mr := "chercher merged_word_with_dash dans toutes les listes"
; IF mr != null →
    keyword_with_dash := true
[] mr = null →
    keyword_with_dash := false
FI

{fusionner les deux mots sans le tiret et chercher si ça forme un mot-clé}
; left_part_without_dash := "récupérer left_part en supprimant le tiret final"
; merged_word_without_dash := left_part_without_dash + right_part
; mr := "chercher merged_word_without_dash dans toutes les listes"
; IF mr != null →
    keyword_without_dash := true
[] mr = null →
    keyword_without_dash := false
FI

{chercher si un des deux mots, pris séparément, forme un mot-clé}
; mr_left := "chercher left_part_without_dash dans toutes les listes"
; mr_right := "chercher right_part dans les listes"
```

```
; IF (mr_left != null) OR (mr_right != null) →
    is_part_a_keyword := true
[] (mr_left = null) AND (mr_right = null) →
    is_part_a_keyword := false
FI

{déterminer merge_attr selon les combinaisons trouvées ci-dessus}
; IF keyword_with_dash = true →
    {les deux mots, avec le tiret, forment un mot-clé}
    merge_attr := "keep-dash"
[] (keyword_with_dash = false) AND
   (keyword_without_dash = true) →
    {les deux mots, sans le tiret, forment un mot-clé}
    merge_attr := "discard-dash"
[] (keyword_with_dash = false) AND
   (keyword_without_dash = false) AND
   (is_part_a_keyword = true) →
    {un des deux mots est un mot-clé}
    merge_attr := "no"
[] (keyword_with_dash = false) AND
   (keyword_without_dash = false) AND
   (is_part_a_keyword = false) →
    {aucun mot-clé n'a été trouvé}
    IF ("left_part_without_dash se termine par un chiffre")
        OR ("right_part commence par un chiffre") →
        merge_attr := "no"
    [] ("left_part_without_dash ne se termine pas par un chiffre")
        AND ("right_part ne commence pas par un chiffre") →
        merge_attr := "discard-dash"
    FI
FI
]|
```

F.2.3 "extraire le dernier mot de left_text"

La commande `"extraire le dernier mot de left_text"` nécessite la déclaration de variables supplémentaires :

```
    delim : char ;
    word : string

|[
{extrait le mot situé juste avant le tiret,
 ou rien s'il y a un délimiteur juste avant le tiret.
 La propriété de délimiteur est donnée dans la table de
 caractères du FEW.}
delim := "chercher dans left_text le premier caractère
          délimiteur à partir de la fin"
; word := "récupérer le texte situé entre delim et la fin de
```

```
left_text"
; IF word in liste_tirets →
    left_part := null
[] word not in liste_tirets →
    left_part := word
FI
]|
```

F.2.4 "extraire le premier mot de right_text"

```
|[
{extrait le mot situé juste après break_tag.
 La propriété de délimiteur est donnée
 dans la table de caractères du FEW.}
delim := "chercher dans right_text le premier caractère délimiteur"
; word := "récupérer le texte situé entre le début de right_text et
delim"
; right_part := word
]|
```

F.3 Split-doc-com

```
|[
VAR
    ck : chunk ;
    n : integer ;
    vol_id, type_attr : string
BEGIN
; "initialiser ck à la balise ouvrante <art>"
; type_attr := "extraire de <art> la valeur de l'attribut type"
; IF type_attr = "renvoi" →
    SKIP
[] (type_attr != "renvoi") OR ("type_attr non défini") →
    n := "compter le nombre de paragraphes de l'article"
    ; IF n = 0 →
        type_attr := "entry-only"
    [] n > 0 →
        vol_id := "extraire de <art> la valeur de l'attribut volume"
        ; IF (vol_id = "21") OR
             (vol_id = "22") OR
             (vol_id = "23") →
            "traiter l'article de concept"
            ; "type_attr := "doc-concept"
        [] (vol_id != "21") AND
           (vol_id != "22") AND
           (vol_id != "23") →
            IF n = 1 →
```

```
                "baliser le paragraphe en tant que <mixt>"
                ; "type_attr := "mixt-only"
            [] n > 1 →
                "choisir la liste de mots-clés"
                ; "traiter l'article"
            FI
        FI
    FI
    ; "ajouter à <art> un attribut type prenant la valeur de
    la variable type_attr"
FI
END
]|
```

F.3.1 "traiter l'article de concept"

La commande "`traiter l'article de concept`" nécessite la déclaration de variables supplémentaires :

```
    contains_geoling : boolean ;
    ck_p : chunk
```

```
|[
contains_geoling := false
; "déplacer ck_p à la première balise <p> après </entry>"
; DO ck_p != null →
    contains_geoling := "chercher balise ouvrante <geoling>
    à partir de ck jusqu'à la balise fermante </p>"
    ; IF contains_geoling = true →
        "baliser le paragraphe en tant que <doc>"
    [] contains_geoling = false →
        "baliser le paragraphe en tant que <com>"
    FI
    ; "déplacer ck_p à la prochaine balise ouvrante <p>"
    ; IF <p> précédé de la balise <notes> →
        ck_p := null
    FI
OD
]|
```

F.3.2 "choisir la liste de mots-clés"

La commande "`choisir la liste de mots-clés`" nécessite la déclaration de variables supplémentaires :

```
    lang : string
```

```
|[
```

```
lang := "extraire de <art> la valeur de l'attribut lang"
; IF lang = "german" →
    "sélectionner la liste few-com-wort-base"
[] lang = "french" →
    "sélectionner la liste few-com-mot-base"
FI
]|
```

F.3.3 "traiter l'article"

La commande "traiter l'article" nécessite la déclaration de variables supplémentaires :

```
    vs : virtualString

|[
"déplacer ck_p à la première balise <p> après </entry>"
; DO ck_p != null →
    vs:= "construire chaîne virtuelle à partir de <p> selon partition()"
    ; "traiter vs"
    ; "déplacer ck_p à la prochaine balise ouvrante <p>"
    ; IF "<p> est précédé de la balise <notes>" →
        ck_p := null
    FI
OD
; "vérifier le séquençage des paragraphes"
]|

partition()
    balises terminales: p
    balises visibles: aucune
    balises transparentes: par défaut
    balises invisibles avec contenu: appelnote, b, biblio, def, i, geoling, étymon
    balises inattendues: aucune
```

F.3.4 "traiter vs"

La commande "traiter vs" nécessite la déclaration de variables supplémentaires :

```
    vs_without_par : virtualString ;
    n : integer ;
    tag_type : tag

|[
vs_without_par := "rendre invisible dans vs le texte entre parenthèses"
; vs_without_par := "version en minuscules de vs_without_par"
; "chercher dans vs_without_par les mots-clés de la liste sélectionnée"
; n := "calculer le nombre de mots-clés trouvés"
```

```
; IF n = 0 →
    tag_type := <doc>
[] n > 0 →
    tag_type := <com>
FI
; "baliser le paragraphe en tant que tag_type"
]|
```

F.3.5 "vérifier le séquençage des paragraphes"

La commande "`vérifier le séquençage des paragraphes`" nécessite la déclaration de variables supplémentaires :

```
    last_p : chunk ;
    last_p_modified : boolean ;
    high_number : boolean ;
    exist_com : boolean

last_p_modified := false

{phase 1}
; "initialiser last_p à la dernière balise <p> de l'article
   non incluse dans un élément <notes>"
; IF "last_p est directement précédé de la balise <doc>" →
    high_num := "calculer si le nombre de <geoling>
                 dans le paragraphe est élevé"
    ; IF high_num = false →
        "changer <doc> en <com> et </doc> en </com>"
        ; last_p_modified := true
    [] high_num = true →
        SKIP
    FI
[] "last_p n'est pas précédé de la balise <doc>" →
    SKIP
FI

{phase 2}
; "initialiser last_doc à la dernière balise <doc> de l'article"
; IF last_doc = null →  {tous les <p> sont dans un élément <com>}
    "changer tous les paragraphes marqués <com> en <mixt>
       sauf le paragraphe commençant à last_p"
    ; type_attr := "mixt-only"
    ; IF last_p_modified = true →
        SKIP
    [] last_p_modified = false →
        high_num := "calculer si le nombre de <geoling> dans le
                     paragraphe commençant par last_p est élevé"
        ; IF high_num = true →
            "marquer le paragraphe commençant par
```

```
            last_p en un paragraphe <mixt>"
        [] high_num = false →
            SKIP
        FI
    FI

{phase 3}
[] last_doc != null →
    exist_com := "chercher une balise <com> avant last_doc"
    ; IF exist_com = true →
        "changer tous les <com> avant last_doc en <mixt>"
    [] exist_com = false →
        SKIP
    FI
    ; "définir l'attribut de <art>"
FI
```

F.3.6 "définir l'attribut de <art>"

```
IF "paragraphes balisés <doc> suivis de paragraphes balisés <com>" →
    type_attr := "doc-com"
[] "uniquement paragraphes balisés <doc>" →
    type_attr := "doc-only"
[] "dans tous les autres cas..."
    type_attr := "doc-mixt"
FI
```

F.4 Split-mixt-art

```
|[
VAR
    ck : chunk ;
    type_attr : string ;
    vs : virtualString
BEGIN
"initialiser ck à la balise ouvrante <art>"
; type_attr := "extraire de <art> la valeur de l'attribut type"
; IF (type_attr = "mixt-only") OR
     (type_attr = "doc-mixt") →
    "vérifier que l'élément <notes> ne contient aucune balise <mixt>"
    ; "sélectionner une liste de mots-clés"
    ; "déplacer ck à la première balise <mixt> après </entry>"
    ; DO ck != null →
        vs := "construire chaîne virtuelle à partir de ck selon partition()"
        ; "traiter vs"
        ; "déplacer ck à la prochaine balise ouvrante <mixt>"
    OD
```

```
[] (type_attr != "mixt-only") AND
     (type_attr != "doc-mixt") →
    SKIP
FI
END
]|

partition()
    balises terminales: p
    balises visibles: aucune
    balises transparentes: par défaut
    balises invisibles avec contenu: appelnote, b, biblio, def, i, geoling
    balises inattendues: aucune
```

F.4.1 "sélectionner une liste de mots-clés"

La commande "`sélectionner une liste de mots-clés`" nécessite la déclaration de variables supplémentaires :

```
    lang : string

|[
lang := "extraire de <art> la valeur de l'attribut lang"
; IF lang = "german" →
    "sélectionner la liste few-com-wort-base"
[] lang = "french" →
    "sélectionner la liste few-com-mot-base"
FI
]|
```

F.4.2 "traiter vs"

La commande "`traiter vs`" nécessite la déclaration de variables supplémentaires :

```
    vs_without_par : virtualString ;
    keyword : string ;
    delim : string ;
    n : integer

|[
vs_without_par := "rendre invisible dans vs le texte entre parenthèses"
; vs_without_par := "version en minuscules de vs_without_par"
; "chercher dans vs_without_par les mots-clés de la liste sélectionnée"
; n := "calculer le nombre de mots-clés trouvés"
; IF n = 0 →
    "alerte et arrêt du traitement"
[] n > 0 →
    keyword := "mot-clé détecté le plus à gauche dans vs_without_par"
```

```
    ; delim := "chercher un caractère délimiteur avant le mot-clé"
    ; IF delim != null →
        "chercher après delim, dans vs_without_par, une balise <geoling>"
        ; IF "pas de balise <geoling>" →
            "baliser les deux champs"
        [] "il y a une balise <geoling>" →
            SKIP
        FI
    [] delim = null →
        SKIP
    FI
FI
]|
```

F.4.3 "chercher un caractère délimiteur avant le mot-clé"

La commande "chercher un caractère délimiteur avant le mot-clé" nécessite la déclaration de variables supplémentaires :

```
    cg : string ;
    pat1, pat2 : regex ;
    num : integer ;
    mr : string
```

```
|[
cg := "extraire contexte à gauche de keyword depuis <p>
       dans vs_without_par"
       {le texte entre parenthèses reste invisible}
; pat1 := /\._?[----]_*/
    {[----] = tiret ou tiret cadratin ou tiret semi-cadratin}
; pat2 := /\./
; num := 0
; mr := "chercher pat1 le plus à droite dans cg"
; IF mr != null →
    delim := mr
    ; num := 1
    ; "mémoriser l'emplacement de delim"
[] mr = null →
    mr := "chercher pat2 le plus à droite dans cg"
    ; IF mr != null →
         delim := mr
        ; num := 2
        ; "mémoriser l'emplacement de delim"
    [] mr = null →
        delim := null
    FI
FI
]|
```

F.4.4 "baliser les deux champs"

La commande "`baliser les deux champs`" nécessite la déclaration d'une variable supplémentaire :

```
    mixt_attr : string
```

```
|[
mixt_attr := "mixt_doc"
"ajouter à la balise <mixt> un attribut mixt-type prenant la
valeur de la variable mixt_attr"
; IF num = 1 →
    "insérer </p></mixt><mixt mixt-type = "mixt_com"><p>
     juste avant delim"
[] num = 2 →
    "insérer </p></mixt><mixt mixt-type = "mixt_com"><p>
     juste après delim"
[] (num != 1) AND (num != 2) →
    "alerte et arrêt du traitement"
FI
]|
```

F.5 Streamline-quotes

```
|[
BEGIN
"détecter les guillemets dans <b>, <e>, <i>, <sc>"
; "normaliser l'article atrium"
; "vérifier l'équilibrage des guillemets, crochets et semi-crochets"
END
]|
```

F.5.1 "détecter les guillemets dans <b>, <e>, <i>, <sc>"

La commande "`détecter les guillemets dans <b>, <e>, <i>, <sc>"`" nécessite la déclaration de variables supplémentaires :

```
    ck : chunk ;
    bal_licites : set of tags ;
    c : char
```

```
|[
bal_licites := {<b>, <e>, <i>, <sc>}
; "initialiser ck à la première balise appartenant à bal_licites"
; DO ck != null →
    vs := "construire chaîne virtuelle à partir de ck
           selon partition(ck)"
```

```
    ; c := "premier caractère de vs"
    ; DO c != null →
        IF (c = U+201C) OR (c = U+201D) OR (c = U+201E) →
            "alerte et arrêt du traitement"
        [] (c != U+201C) AND (c != U+201D) AND (c != U+201E) →
            SKIP
        FI
        ; c:= "caractère suivant dans vs"
        OD
    ; "déplacer ck à la prochaine balise appartenant à bal_licites"
OD
]|

partition(tag ck)
    balises terminales: ck
    balises visibles: aucune
    balises transparentes: par défaut
    balises invisibles: aucune
    balises inattendues: aucune
```

F.5.2 "normaliser l'article ATRIUM"

La commande "normaliser l'article ATRIUM" nécessite la déclaration de variables supplémentaires :

```
    vol_id : integer ;
    first_page_id : integer ;
    vs : virtualString

|[
vol_id := "extraire de <art> la valeur de l'attribut volume"
; first_page_id := "extraire le numéro de page du début de l'article"
; IF (vol_id != "25") OR (first_page_id != "687") →
    SKIP
  []  (vol_id = "25") AND (first_page_id = "687") →
    first_page_id := "689"
    "déplacer ck à la balise <col> avec un attribut pg ayant
     la valeur de la variable first_page_id"
    ; c := "premier caractère situé juste après ck"
    ; vs := "construire chaîne virtuelle à partir de c selon
             partition-fix-guill()"
    ; DO c != null →
        IF c = U+201C →
            c := U+201E
            "remplacer dans vs le caractère correspondant à c
            par U+201E"
        [] c = U+201D →
            c := U+201C
        [] (c != U+201C) AND (c = U+201D) →
```

```
            SKIP
        FI
        ; c := "caractère suivant dans vs"
    OD
FI
]|

partition-fix-guill()
    balises terminales: <art>
    balises transparentes: par défaut
    balises visibles: aucune
    balises invisibles: aucune
    balises inattendues: aucune
```

F.5.3 "vérifier l'équilibrage des guillemets, crochets et semi-crochets"

```
|[
"définir les guillemets licites"
; "déplacer ck à la balise ouvrante <h>"
; vs := "construire chaîne virtuelle à partir de <h> selon partition(h)"
; "traiter vs"
; "déplacer ck à la première balise ouvrante <p>"
; DO ck != null →
    vs := "construire chaîne virtuelle à partir de <p> selon partition(p)"
    ; "traiter vs"
    ; "déplacer ck à la prochaine balise ouvrante <p>"
OD
]|

partition(tag ck)
    balises terminales: ck
    balises visibles: aucune
    balises transparentes: par défaut
    balises invisibles: aucune
    balises inattendues: aucune
```

F.5.4 "définir les guillemets licites"

La commande "`définir les guillemets licites`" nécessite la déclaration de variables supplémentaires :

```
    opening_def_quote, closing_def_quote, illicit_def_quote : char ;
    liste_illicit_quotes : set of char ;
    liste_opening_quotes : set of char ;
    liste_closing_quotes : set of char

|[
IF (vol_id = "25") AND (first_page_id = "689") →
```

```
    opening_def_quote := U+201C
    ; closing_def_quote := U+201D
    ; illicit_def_quote := U+201E
[] (vol_id = "25") AND (first_page_id = "688") →
    "alerte"
[] (vol_id != "25") OR
    ((first_page_id != "688") AND (first_page_id = "689")) →
    opening_def_quote := U+201E
    ; closing_def_quote := U+201C
    ; illicit_def_quote := U+201D
FI
; liste_illicit_quotes := [illicit_def_quote]
; liste_opening_quotes := ["[", opening_def_quote, "U+231C"]
; liste_closing_quotes := ["]", closing_def_quote, "U+231D"]
]|
```

F.5.5 "traiter vs"

La commande "`traiter vs`" nécessite la déclaration de variables supplémentaires :

```
    opening_c : char ;
    x, y : integer

|[
; "créer pile"
; char c := "aller au premier caractère de vs"
; DO c != null →
    IF c in liste_illicit_quotes →
        "émettre avertissement"
    [] c in liste_opening_quotes →
        "insérer c au sommet de la pile"
    [] c in liste_closing_quotes →
        IF "il n'y a rien dans la pile" →
            "émettre alerte"
        [] "il y a quelque chose dans la pile"
            opening_c := "récupérer l'élément au sommet de la pile"
            {vérifier les paires}
            ; x := "numéro d'index de opening_c dans liste_opening_quotes"
            ; y := "numéro d'index de c dans liste_closing_quotes"
            ; IF x = y →
                SKIP
              [] x != y →
                "émettre alerte"
            FI
        FI
    [] (c not in liste_illicit_quotes) AND
       (c not in liste_opening_quotes) AND
       (c not in liste_closing_quotes) →
        SKIP
```

```
    FI
    ; c := "aller au caractère suivant dans vs"
OD
; IF "la pile n'est pas vide" →
    "émettre alerte"
FI
]|
```

F.6 Tag-affix

```
|[
VAR
    ck : chunk
    p_licites : set of tags
BEGIN
"initialiser ck à la première balise <p> après </entry>"
; DO ck != null →
    "baliser les suffixes étymologiques connus"
    ; "baliser les préfixes étymologiques connus"
    ; "déplacer ck à la prochaine balise ouvrante <p>"
OD
p_licites : [<com>, <mixt_com>]
; "déplacer ck à la première balise de l'article appartenant à p_licites"
; DO ck != null →
    "baliser les affixes non répertoriés"
    ; "déplacer ck à la prochaine balise appartenant à p_licites"
OD
END
]|
```

F.6.1 "baliser les suffixes étymologiques connus"

La commande "baliser les suffixes étymologiques connus" nécessite la déclaration de variables supplémentaires :

```
    vs : virtualString ;
    keyword : string ;
    attr_type : string
```

```
|[
vs := "construire chaîne virtuelle à partir de <p> selon partition"
; "chercher dans vs tous les mots-clés de few-suffix-base"
; keyword := "premier mot-clé détecté dans vs"
; attr_type := "suffix"
; DO "keyword != null" →
    "insérer, juste avant keyword, une balise <affix> avec
     un attribut type prenant la valeur de la variable attr_type"
```

```
    ; keyword := "mot-clé suivant détecté dans vs"
OD
]|
```

F.6.2 "baliser les préfixes étymologiques connus"

La commande "baliser les préfixes étymologiques connus" nécessite la déclaration de variables supplémentaires :

```
    licit_prefix : boolean
```

```
|[
licit_prefix := false
vs := "construire chaîne virtuelle à partir de <p> selon partition"
; "chercher dans vs tous les mots-clés de few-prefix-base"
; keyword := "premier mot-clé détecté dans vs"
; DO "keyword != null" →
    licit_prefix := "vérifier la valeur du tiret final"
    ; IF licit_prefix = true →
        ; "insérer, juste avant keyword, une balise <affix> avec
           un attribut type prenant la valeur de la variable attr_type"
        ; "insérer, juste après keyword, une balise </affix>"
    [] licit_prefix = false →
        SKIP
    FI
OD
]|
```

F.6.3 "vérifier la valeur du tiret final"

```
|[
IF ("keyword est suivi de <lb/>") OR
     ("keyword est suivi de <col/>") →
    licit_prefix := false
[] ("keyword n'est pas suivi de <lb/>") OR
     ("keyword n'est pas suivi de <col/>") →
    licit_prefix := true
FI
]|
```

```
partition
    balises terminales: p
    balises visibles: aucune
    balises transparentes: col, lb; par défaut
    balises invisibles avec contenu: e, appelnote, pnum, pref, def, renvoi, affix
    balises inattendues: aucune
```

F.6.4 "baliser les affixes non répertoriés"

La commande "`baliser les affixes non répertoriés`" nécessite la déclaration de variables supplémentaires :

```
    ck_i : chunk

"initialiser ck_i à la première balise <i>"
DO ck_i != null →
    "construire vs à partir de <i> selon partition-affix()"
    ; IF  (matche "-" juste après <i> dans vs) AND
        (matche "-" juste avant </i> dans vs) →
        attr_type := "?"
        ; "baliser vs avec un attribut
      type prenant la valeur de la variable attr_type"
    [] (matche "-" juste après <i> dans vs) AND
    (ne matche pas "-" juste avant </i> dans vs) →
        attr_type := "suffix"
        ; "baliser vs avec un attribut
      type prenant la valeur de la variable attr_type"
    [] (ne matche pas "-" juste après <i> dans vs) AND
        (matche "-" juste avant </i> dans vs) →
        attr_type := "prefix"
        ; "baliser vs avec un attribut
      type prenant la valeur de la variable attr_type"
    [] (ne matche pas "-" juste après <i> dans vs) AND
        (ne matche pas "-" juste avant </i> dans vs) →
        SKIP
    FI
    ; "déplacer ck_i à la prochaine balise <i>"
OD

partition-affix()
    balises terminales: <i>
    balises visibles: aucune
    balises transparentes: <lb/>, <col/>
    balises invisibles: <appelnote>, <pnum>, <pref>, <e>
    balises inattendues: aucune
```

F.7 Tag-appelnote

```
|[
VAR
    ck : chunk ;
    nbr_notes : integer
BEGIN
"initialiser ck à la balise ouvrante <notes>"
; IF ck = null →
```

```
    SKIP
[] ck != null →
    nbr_notes := "extraire de <notes> la valeur de l'attribut count"
    ; "traiter l'entrée"
    ; "déplacer ck à la première balise ouvrante <p> après </entry>"
    ; DO ck != null →
        ; "traiter le paragraphe"
        ; "déplacer ck à la prochaine balise ouvrante <p>"
    OD
    ; "vérifier la correspondance entre notes et appels de note"
FI
END
]|
```

F.7.1 "traiter l'entrée"

La commande `"traiter l'entrée"` nécessite la déclaration de variables supplémentaires :

```
    ck_e : chunk ;
    vs : virtualString ;
    deux_appels : boolean

|[
"initialiser ck_e à la première balise ouvrante <e>"
; DO ck_e != null →
    vs := "construire chaîne virtuelle à partir de <e> selon partition-entry()"
    ; deux_appels := false
    ; "traiter vs"
    ; IF deux_appels = true →
        "déplacer ck_e au chunk situé juste après </appelnote>"
    [] deux_appels = false →
        "déplacer ck_e au chunk situé juste après </e>"
    FI
    ; "déplacer ck_e à la prochaine balise <e>"
OD
]|

partition-entry()
    balises terminales: h, e
    balises visibles: aucune
    balises transparentes: par défaut
    balises invisibles avec contenu: aucune
    balises inattendues: aucune
```

F.7.2 "traiter le paragraphe"

```
|[
```

```
"déplacer ck_e à la première balise ouvrante <e> avant </p>"
; DO ck_e != null →
    vs := "construire chaîne virtuelle à partir de <e> selon partition-p"
    ; deux_appels := false
    ; "traiter vs"
    ; IF deux_appels = true →
        "déplacer ck_e au chunk situé juste après </appelnote>"
    [] deux_appels = false →
        "déplacer ck_e au chunk situé juste après </e>"
    FI
    ; "déplacer ck_e à la prochaine balise <e> avant </p>"
OD
]|

partition-p
Balises terminales: p, e
Balises visibles: aucune
Balises transparentes: par défaut
Balises invisibles avec contenu: aucune
Balises inattendues: aucune
```

F.7.3 "traiter vs"

La commande "traiter vs" nécessite la déclaration de variables supplémentaires :

```
VAR
    pat : regex ;
    mr : string ;
    note_id : integer ;
    status_note_id : boolean

|[
pat := /\\s*[0-9]+\\s*\\)?/
; mr := "matcher regex dans vs"
; IF mr = null →
    SKIP
[] mr != null →
    note_id := "extraire de mr le numéro de note"
    ; status_note_id := true
    ; IF (note_id > note_count) OR (note_id < 1) →
        status_note_id := false
    FI
    ; IF "il y a une parenthèse fermante dans <e>...</e>" →
        "baliser l'appel de note"
        ; "mémoriser note_id et status_note_id"
    [] "il n'y a pas de parenthèse fermante dans <e>...</e>" →
        IF "il y a une parenthèse fermante juste après </e>" →
            "vérifier la succession des parenthèses, baliser si ok"
        [] "il n'y a pas de parenthèse fermante juste après </e>" →
```

```
            SKIP
        FI
    FI
FI
]|
```

F.7.4 "vérifier la succession des parenthèses, baliser si ok"

La commande "`vérifier la succession des parenthèses, baliser si ok`" nécessite la déclaration de variables supplémentaires :

```
    vs_gauche : virtualString
    parenthesis : string
```

```
|[
vs_gauche := "construire chaîne virtuelle à gauche de
              mr, depuis <p>, selon partition-e"
; IF vs_gauche = null →
    "alerte et arrêt du traitement"
[] vs_gauche != null →
    SKIP
FI
; parenthesis := "chercher dans vs_gauche la parenthèse la plus
                 proche de mr (donc la plus à droite dans vs_gauche)"
; IF (parenthesis = null) OR (parenthesis = ")") →
    "baliser l'appel de note"
    ; "mémoriser note_id et status_note_id"
[] (parenthesis != null) AND (parenthesis = "(") →
    {on est dans le cas "(...X)"}
    "examiner le contexte droit, baliser si ok"
FI
]|
```

```
partition-e
    balises terminales: p
    balises visibles: aucune
    balises transparentes: par défaut
    balises invisibles: e
    balises inattendues: aucune
```

F.7.5 "examiner le contexte droit, baliser si ok"

La commande "`examiner le contexte droit, baliser si ok`" nécessite la déclaration de variables supplémentaires :

```
    vs_droite : virtualString ;
    second_appel : boolean
```

```
|[
vs_droite := "construire chaîne virtuelle à partir de mr
              selon partition-e"
; IF vs_droite = null →
    SKIP
[] vs_droite != null →
    parenthesis := "chercher dans vs_droite la parenthèse la plus
                    proche de mr (donc la plus à gauche dans vs_droite)"
    ; IF (parenthesis = null) OR (parenthesis = "(") →
        SKIP   {pas de balisage}
    [] (parenthesis = null) AND (parenthesis = ")") →
        {on est dans le cas "(...X)...)"}
        second_appel := "chercher un chiffre en exposant devant parenthesis"
        ; IF second_appel = false →
            "baliser l'appel de note"
            ; "mémoriser note_id et status_note_id"
        [] second_appel = true →    {on est dans le cas "(...X)...Y)"}
            "traiter les deux appels, baliser en marquant ambigu"
            ; deux_appels := true
        FI
    FI
FI
]|
```

F.7.6 "traiter les deux appels, baliser en marquant ambigu"

La commande "`traiter les deux appels, baliser en marquant ambigu`" nécessite la déclaration de variables supplémentaires :

```
    second_note_id : integer ;
    status_second_note_id : boolean
```

```
|[
second_note_id := "extraire le numéro de note du second appel"
; status_second_note_id := true
; IF (second_note_id > note_count) OR (second_note_id < 1) →
    status_second_note_id := false
FI
; IF status_note_id = status_second_note_id →
    "baliser les deux appels de note en les marquant ambigus"
    ; "mémoriser notes_id et status_note_id"
    ; "mémoriser second_note_id et status_second_note_id"
[] status_note_id != status_second_note_id →
    "baliser les deux appels de note en marquant ambigu
     celui dont le statut est false"
    ; "mémoriser notes_id et status_note_id"
    ; "mémoriser second_note_id et status_second_note_id"
FI
]|
```

F.7.7 "vérifier la correspondance entre notes et appels de note"

La commande "vérifier la correspondance entre notes et appels de note" nécessite la déclaration de variables supplémentaires :

```
    ordre_croissant : boolean ;
    missing_notes, repeated_notes : integer ;
    value : string
```

```
|[
"extraire tous les note_id et status_note_id
 dans l'ordre dans lequel ils ont été mémorisés"
; ordre_croissant := "vérifier l'ordre croissant des note_id dont
                      le statut est true"
; missing_notes := "compter le nombre d'appels de note manquants"
; repeated_notes := "compter le nombre de note_id redondants"
; value := "définir la valeur de l'attribut call-sequence"
; "ajouter à la balise <notes> un attribut call-sequence
   prenant pour valeur le contenu de la variable value"
]|
```

F.7.8 "définir la valeur de l'attribut call-sequence"

```
|[
value := null
; IF missing_notes > 0 →
    value := "incomplete"
[] (missing_notes = 0) AND (ordre_croissant = false) →
    value := "non-monotonic"
[] (missing_notes = 0) AND (ordre_croissant = true) AND
   (repeated_notes > 0) →
    value := "some-redundancy"
[] (missing_notes = 0) AND (ordre_croissant = true) AND
   (repeated_notes = 0) →
    value := "ok"
FI
]|
```

F.8 Tag-attestation

```
|[
VAR
    ck, ck_precisions : chunk ;
    p_licites : set of tags ;
    attr_status : string
BEGIN
p_licites := [<doc>, <mixt>]
```

```
; "initialiser ck à la première balise de l'article appartenant à p_licites"
; DO ck != null →
    "initialiser ck_precisions à la première balise ouvrante
    <precisions> du paragraphe"
    ; DO ck_precisions != null →
        attr_status := "extraire de <precisions> la valeur de l'attribut status"
        ; IF (attr_status = "ok-wrap-parentheses") OR
             (attr_status = "ambiguous") →
             "baliser les attestations"
        [] (attr_status != "ok-wrap-parentheses") AND
             (attr_status != "ambiguous") →
             SKIP
        FI
        ; "déplacer ck_precisions à la prochaine balise ouvrante
           <precisions> du paragraphe"
    OD
    ; "déplacer ck à la prochaine balise de l'article appartenant à p_licites"
OD
END
]|
```

F.8.1 "baliser les attestations"

La commande "baliser les attestations" nécessite la déclaration de variables supplémentaires :

```
    vs : virtualString ;
    mr : string ;
    pat : regex

|[
vs := "construire chaîne virtuelle à partir de <precisions> selon partition()"
; "insérer <attestation> au début de vs, juste après la parenthèse ouvrante"
; pat := /;/
; mr := "matcher pat dans vs"
; DO mr != null →
    "insérer la balise </attestation> juste avant mr"
    ; "insérer la balise <attestation> juste après mr"
    ; mr := "matcher pat suivant dans vs"
OD
; "insérer </attestation> à la fin de vs, juste avant la parenthèse fermante"
]|

partition()
    balises terminales: precisions
    balises visibles: aucune
    balises transparentes: aucune
    balises invisibles: par défaut
    balises inattendues: aucune
```

F.9 Tag-biblio

```
|[
VAR
    ck : chunk ;
    vs : virtualString
BEGIN
"créer la liste de collisions bib-geoling"
; "initialiser ck à la première balise <p> après </entry>"
; DO ck != null →
    vs := "construire chaîne virtuelle à partir de <p> selon partition()"
    ; "traiter vs"
    ; "déplacer ck à la prochaine balise ouvrante <p>"
OD
END
]|

partition()
    balises terminales: p
    balises visibles: aucune
    balises transparentes: par défaut
    balises invisibles: appelnote, e, def, i, lang_etymon, renvoi, signature
    balises inattendues: entry
```

F.9.1 "créer la liste de collisions bib-geoling"

La commande `"créer la liste de collisions bib-geoling"` nécessite la déclaration de variables supplémentaires :

```
    colliding_base : set of string ;
    keyword : string
```

```
|[
colliding_base := []
; keyword := "premier mot-clé de few-bib-base"
; DO keyword != null →
    IF "keyword appartient à few-geoling-base" →
        "ajouter keyword à colliding-base"
    [] "keyword n'appartient pas à few-geoling-base" →
        SKIP
    FI
    ; IF "keyword appartient à few-geoling-error-base" →
        "ajouter keyword à colliding-base"
    [] "keyword n'appartient pas à few-geoling-error-base" →
        SKIP
    FI
    ; IF "keyword se termine par un point" →
        keyword := "supprimer le point final de keyword"
```

```
        ; IF "keyword appartient à few-geoling-base" →
            "ajouter keyword à colliding-base"
        [] "keyword n'appartient pas à few-geoling-base" →
            SKIP
        FI
        ; IF "keyword appartient à few-geoling-error-base" →
            "ajouter keyword à colliding-base"
        [] "keyword n'appartient pas à few-geoling-error-base" →
            SKIP
        FI
    [] "keyword ne se termine pas par un point" →
        keyword := "ajouter un point à la fin de keyword"
        ; IF "keyword appartient à few-geoling-base" →
            "ajouter keyword à colliding-base"
        [] "keyword n'appartient pas à few-geoling-base" →
            SKIP
        FI
        ; IF "keyword appartient à few-geoling-error-base" →
            "ajouter keyword à colliding-base"
        [] "keyword n'appartient pas à few-geoling-error-base" →
            SKIP
        FI
    FI
    ; keyword := "mot-clé suivant de few-bib-base"
OD
]|
```

F.9.2 "traiter vs"

La commande "`traiter vs`" nécessite la déclaration d'une variable supplémentaire :

```
    status_attr : string

|[
"chercher dans vs tous les mots-clés de few-bib-base"
; keyword := "premier mot-clé trouvé dans vs"
; status_attr := "ok"
; DO keyword != null →
    IF "keyword appartient à colliding-base" →
        status_attr := "geoling?"
        ; "émettre avertissement"
    [] "keyword n'appartient pas à colliding-base" →
        SKIP
    FI
    ; "baliser keyword en tant que <biblio> avec un attribut status prenant la
       valeur de la variable status_attr"
    ; "inclure les références attenantes"
    ; keyword := "mot-clé suivant trouvé dans vs"
OD
```

```
]|
```

F.9.3 "inclure références attenantes"

La commande "`inclure références attenantes`" nécessite la déclaration de variables supplémentaires :

```
    number, pat1, pat2, pat3, pat : regex
    mr : string
```

```
|[
number := /([1-9][0-9]*)/
; pat1 := /␣?/ + number + /(,?␣/ + number + /(␣?[----]␣?/
          + number + /)?)*/
; pat2 := /␣?/ + number + /\*?(␣?p␣?/ + number + /)?(,␣?/
          + number + /)*/
; pat3 := /␣?/ + number + /(␣?[----]␣?/ + number + /)?/
; pat := /([/ + pat1 + pat2 + pat3 + /](␣?;␣?[/ + pat1 + pat2
          + pat3 + /])*)?/
; mr := "matcher pat juste après </biblio>"
; IF mr = null →
    SKIP
[] mr != null →
    "déplacer la balise </biblio> juste après mr"
FI
]|
```

F.10 Tag-concept

```
|[
VAR
    vol_id : integer;
BEGIN
vol_id := "extraire de <art> la valeur de l'attribut volume"
; IF (vol_id = "21") OR (vol_id = "22") OR (vol_id = "23") →
    "traiter l'entrée"
    ; "traiter les paragraphes"
[] (vol_id != "21") AND (vol_id != "22") AND (vol_id != "23") →
    "traiter les paragraphes"
FI
END
]|
```

F.10.1 "traiter l'entrée"

La commande "`traiter l'entrée`" nécessite la déclaration de variables supplémentaires :

```
    vs : virtualString ;
    n : integer

|[
vs := "construire une chaîne virtuelle à partir de <entry>
        selon partition-entry()"
; "chercher les mots-clés de few-concept-base dans vs"
; n := "compter le nombre de mots-clés détectés dans vs"
; IF (n = 0) OR (n > 1) →
    "émettre un avertissement et stopper le traitement"
[] n = 1 →
    "vérifier que le mot-clé est inclus dans des balises <b>...</b>"
    ; IF "le test est positif" →
        "baliser le mot-clé en intégrant les balises <b>...</b>"
    [] "le test est négatif" →
        "baliser le mot-clé"
        ; "émettre un avertissement"
    FI
FI
]|

partition-entry()
    balises terminales: entry
    balises visibles: aucune
    balises transparentes: par défaut
    balises invisibles avec contenu: aucune
    balises inattendues: etymon
```

F.10.2 "traiter les paragraphes"

La commande `"traiter les paragraphes"` nécessite la déclaration d'une variable supplémentaire :

```
    ck : chunk

|[
"initialiser ck à la première balise <p> après </entry>"
; DO ck != null →
    vs := "construire chaîne virtuelle à partir de <p> selon partition-p()"
    ; "traiter vs"
    ; "déplacer ck à la prochaine balise ouvrante <p>"
OD
]|

partition-p()
    balises terminales: p
    balises visibles: aucune
    balises transparentes: biblio; par défaut
    balises invisibles avec contenu: etymon, appelnote, e, i, sc
```

```
balises inattendues: aucune
```

F.10.3 "traiter vs"

La commande "`traiter vs`" nécessite la déclaration d'une variable supplémentaire :

```
    keyword : string

|[
"chercher dans vs tous les mots-clés de few-concept-base"
; keyword := "premier mot-clé détecté dans vs"
; DO keyword != null →
    "vérifier que keyword est précédé d'une flèche (U+2192)"
    IF "test positif" →
        "baliser le mot-clé en tant que <concept>"
        ; "baliser le mot-clé en tant que <renvoi> en intégrant la flèche"
        ; IF keyword = "Vierge" →
            "supprimer balises <biblio> éventuellement situées autour de keyword"
        [] keyword != "Vierge" →
            SKIP
        FI
    [] "test négatif" →
        SKIP
    FI
    ; keyword := "mot-clé suivant détecté dans vs"
OD
]|
```

F.11 Tag-date

```
|[
VAR
    ck : chunk ;
    attr_type : string
BEGIN
attr_type := "extraire la valeur de l'attribut type de <art>"
; IF attr_type = "renvoi" →
    SKIP
[] attr_type != "renvoi" →
    "définir les expressions régulières selon métalangue"
    ; "initialiser ck à la première balise <p> après </entry>"
    ; DO ck != null →
        "traiter le paragraphe"
        ; "déplacer ck à la prochaine balise <p>"
    OD
END
]|
```

```
partition()
    balises terminales: p
    balises visibles: aucune
    balises transparentes: par défaut
    balises invisibles: biblio, renvoi, date, appelnote
    balises inattendues: aucune
```

F.11.1 "définir les expressions régulières selon métalangue"

La commande "`définir les expressions régulières selon métalangue`" nécessite la déclaration de variables supplémentaires :

```
    pat1, pat2, pat3, pat4 : regex ;
    pat2_german, pat3_german, pat4_german : regex ;
    pat2_french, pat3_french, pat4_french : regex ;
    lang : string

|[
pat1 := /(ca\._?|env\._?)?(1[0-9]{3})/
    {date entre 1000 et 1999, éventuellement précédée de "ca." ou "env."}
; pat2_german := /(9|1[0-9]|20)\._?jh\./
    {ex. "13. jh.": nombre de 9 à 20 + point + espace? + "jh."}
; pat3_german := /((9|1[0-9]|20)\.)_?[----]_?((9|1[0-9]|20)\._?jh\.)/
    {ex. "13.-15. jh."}
    {on balise séparément "13." et "15. jh."}
; pat4_german := /([1234]_?\._?viertel|[123]_?\._?drittel|[12]_?\.
                  _?hälfte|anfang|mitte|ende)_?((9|1[0-9]|2[0-1])\._?jh\.)/
; pat2_french := /((9)|(1[0-9])|(20))e_?s\./
    {nombre de 9 à 20 + "e" en exposant + espace? + "s."}
; pat3_french := /((9|1[0-9]|20)e)_?[----]_?((9|1[0-9]|20)e_?s\.)/
    {ex. "13e-15e s."}
    {on balise séparément "13e" et "15e s."}
; pat4_french := /((1er|2e|3e)_?t\._?|
                  (1ère|2e)_?m\._?|
                  (1er|[234]e)_?q\._?|
                  mil\._?|déb\._?|fin_?)
                  (9|1[0-9]|20)e_?s\./
    {ex. "1er t. 12e s."}

; lang := "extraire de <art> la valeur de l'attribut lang"
; IF lang = "french" →
    pat2 := pat2_french
    ; pat3 := pat3_french
    ; pat4 := pat4_french
[] lang= "german" →
    pat2 := pat2_german
    ; pat3 := pat3_german
    ; pat4 := pat4_german
```

```
[] (lang != "french") AND (lang != "german") →
    "alerte et arrêt du traitement"
FI
|]
```

F.11.2 "traiter le paragraphe"

La commande `"traiter le paragraphe"` nécessite la déclaration d'une variable supplémentaire :

```
    mr : string
```

```
vs := "construire chaîne virtuelle à partir de <p> selon partition()"
; mr := "matcher pat1 dans vs"
; DO mr != null →
    "baliser mr en tant que <date>"
    ; mr := "matcher pat1 dans vs"
OD
; vs := "construire vs à partir de <p> selon partition()"
; mr := "matcher pat4 dans vs"
; DO mr != null →
    "baliser mr en tant que <date>"
    ; mr := "matcher pat4 dans vs"
OD
; vs := "construire vs à partir de <p> selon partition()"
; mr := "matcher pat3 dans vs"
; DO mr != null →
    "baliser la section couverte par $1 en tant que <date>"
    ; "baliser la section couverte par $2 en tant que <date>"
    ; mr := "matcher pat3 dans vs"
OD
; vs := "construire vs à partir de <p> selon partition"
; mr := "matcher pat2 dans vs"
; DO mr != null →
    "baliser mr en tant que <date>"
    ; mr := "matcher pat2 dans vs"
OD
```

F.12 Tag-def

```
|[
VAR
    ck : chunk ;
    vs : virtualString ;
    zone_def : boolean ;
    c : char ;
    idx, debut_zone : integer ;
```

```
    guill_ouvrant, guill_fermant, guill_interdit : char
BEGIN
"définir guillemets"
; "initialiser ck à la première balise ouvrante <p> après </entry>"
; DO ck != null →
    vs := "construire chaîne virtuelle à partir de <p> selon partition()"
    ; idx := 0
    ; zone_def := false
    ; c := "premier caractère dans vs"
    ; DO c != null →
        IF c = guill_ouvrant →
            IF zone_def = true →
                "alerte et arrêt du traitement"
            [] zone_def = false →
                zone_def := true
                ; debut_zone := idx
            FI
        [] c = guill_fermant →
            IF zone_def = false →
                alerte et arrêt du traitement
            []  zone_def = true →
                zone_def := false
                ; "baliser en tant que <def> la zone depuis debut_zone
                   caractères après le début de vs, jusque c inclus"
            FI
        [] c = guill_interdit →
            "alerte et arrêt du traitement"
        [] (c != guill_ouvrant) AND (c != guill_fermant)
           AND (c != guill_interdit) →
            SKIP
        FI
        ; c := "caractère suivant dans vs"
        ; idx := idx + 1
    OD
    ; "déplacer ck à la prochaine balise ouvrante <p>"
OD
; IF zone_def = true →
    "alerte et arrêt du traitement"
[] zone_def = true →
    SKIP
FI
END
]|

partition()
balises terminales: p
balises visibles: aucune
balises transparentes: par défaut
balises invisibles: b, e, i, sc
balises inattendues: entry
```

F.12.1 "définir guillemets"

La commande "`définir guillemets`" nécessite la déclaration de variables supplémentaires :

```
    vol_id, pg_id : integer
```

```
|[
vol_id := "extraire le numéro de volume de l'article"
; pg_id := "extraire le numéro de page de l'article"
; IF (vol_id = 25) AND (pg_id >= 687) →
    guill_ouvrant := U+201C
    ; guill_fermant := U+201D
    ; guill_interdit := U+201E
[] (vol_id != 25) OR (pg_id < 687) →
    guill_ouvrant := U+201E
    guill_fermant := U+201C
    guill_interdit := U+201D
FI
]|
```

F.13 Tag-etymon

```
|[
VAR
    ved : string ;
    ck : chunk
BEGIN
"initialiser ck à la balise ouvrante <entry>"
; "traiter les étymons dans l'entrée"
; ved := "extraire dans <entry> le contenu de la première balise <etymon>
          dont l'attribut type a "vedette" pour valeur>"
; "déplacer ck à la première balise ouvrante <p> de l'article"
; DO ck != null →
    "traiter les étymons dans le paragraphe"
    ; "déplacer ck à la prochaine balise <p> de l'article"
OD
END
]|
```

F.13.1 "traiter les étymons dans l'entrée"

La commande "`traiter les étymons dans l'entrée`" nécessite la déclaration de variables supplémentaires :

```
    n : integer ;
    attr_type : string ;
```

```
    attr_desc : string ;
    keyword : string

|[
; vs := "construire chaîne virtuelle à partir de <entry> selon partition()"
; "chercher tous les mots-clés de few-etymon-base dans vs"
; n := "calculer le nombre d'occurrences trouvées"
; IF n = 0 →
    erreur ("aucun étymon trouvé dans <entry>")
[] n > 3 →
    "alerte et arrêt du traitement"
[] n = 1 →
    "définir bal"
    ; attr_type := "définir type en fonction de bal"
    ; IF attr_type = "vedette" →
        "baliser le mot-clé <etymon> avec un attribut type prenant
        la valeur de attr_type"
    [] attr_type != "vedette" →
        "alerte"
    FI
[] (n = 2) OR (n = 3) →
    keyword := "premier mot-clé détecté dans vs"
    ; DO keyword != null →
        "définir bal"
        ; attr_type = "définir type en fonction de bal"
        ; "baliser keyword <etymon> avec un attribut type prenant
        la valeur de attr_type"
        ; keyword := "mot-clé suivant détecté dans vs"
    OD

    {vérifier la validité du balisage en fonction des attributs}
    ; IF (type du 1er <etymon> = "vedette")
        AND
        (type du 2e <etymon> = "vedette" ou "renvoi")
        AND
        ((pas de 3e <etymon>) OR
         (type du 3e <etymon> = "renvoi")) →
        SKIP
    []  (attribut du 1er <etymon> != "vedette")
        OR
        (attribut du 2e <etymon> != "vedette" ou "renvoi")
        OR
        ((il existe un 3e <etymon>) AND
         (type du 3e <etymon> != "renvoi")) →
        "alerte et arrêt du traitement"
    FI
FI
]|

partition()
```

```
balises terminales: entry
balises visibles: aucune
balises transparentes: b, i, sc, e, lang_etymon
balises invisibles: appelnote
balises inattendues: p
```

F.13.2 "définir bal"

La commande "`définir bal`" nécessite la déclaration de variables supplémentaires :

```
clt, ctr, bal : tag
```

```
|[
clt := "déterminer la balise <b>, <i> ou <sc> ouvrante
        la plus proche vers la gauche du premier caractère du mot-clé
        (à l'exclusion des balises équilibrées)"
; crt := "déterminer la balise </b>, </i> ou </sc> fermante
        la plus proche vers la droite du dernier caractère du mot-clé
        (à l'exclusion des balises équilibrées)"
; IF "clt correspond à crt" →
    bal := clt
[] "clt ne correspond pas à crt" →
    IF (clt se trouve juste à gauche du mot-clé) AND
        (crt se trouve juste à droite du mot-clé) →
        "alerte et arrêt du traitement"
    [] (clt ne se trouve pas juste à gauche du mot-clé) AND
        (crt ne se trouve pas juste à droite du mot-clé) →
        "alerte et arrêt du traitement"
    [] (clt se trouve juste à gauche du mot-clé) AND
        (crt ne se trouve pas juste à droite du mot-clé) →
        bal := clt
    [] (clt ne se trouve pas juste à gauche du mot-clé) AND
        (crt se trouve juste à droite du mot-clé) →
        bal := crt
    FI
FI
]|
```

F.13.3 "définir type en fonction de bal"

```
|[
IF "traitement de l'entrée" →
    IF (bal = <b>) OR (bal = </b>) →
        attr_type := "vedette"
    [] (bal = <sc>) OR (bal = </sc>) →
        attr_type := "renvoi"
    [] (bal = <i>) OR (bal = </i>) →
        attr_type := "?"
```

```
    FI
[] "traitement d'un paragraphe" →
    IF (bal = <sc>) OR (bal = </sc>) →
        attr_type := "renvoi"
    [] (bal = <i>) OR (bal = </i>) →
        attr_type := "?"
    [] (bal = <b>) OR (bal = </b>) →
        attr_type := "?"
    FI
FI
]|
```

F.13.4 "définir desc en fonction de bal"

```
|[
IF (bal = <sc>) OR (bal = </sc>) →
    attr_desc := "hereditaire"
[] (bal = <i>) OR (bal = </i>) →
    attr_desc := "emprunts"
[] (bal = <b>) OR (bal = </b>) →
    attr_desc := "?"
FI
]|
```

F.13.5 "traiter les étymons dans le paragraphe"

```
|[
{balisage des sous-lemmes (étymons cachés)}
vs := "construire chaîne virtuelle à partir de <p>
       selon partition-sous-lemmes()"
; "chercher tous les mots-clés de few-etymon-cache-base dans vs"
; keyword := "premier mot-clé détecté dans vs"
; DO keyword != null →
    attr_type := "sous-lemme"
    ; "baliser keyword <etymon> avec un attribut type prenant
    la valeur de la variable attr_type"
    ; keyword := "mot-clé suivant détecté dans vs"
OD

{balisage des lemmes}
; vs := "construire chaîne virtuelle à partir de <p>
        selon partition-lemmes()"
; "chercher tous les mots-clés de few-etymon-base dans vs"
; keyword := "premier mot-clé détecté dans vs"
; DO keyword != null →
    IF keyword = ved → {cf. bloc principal}
        attr_type := "vedette"
        ; "définir bal"
```

```
        ; attr_desc := "définir desc en fonction de bal"
        ; "baliser keyword <etymon> avec les attributs type et desc prenant
        la valeur des variables attr_type et attr_desc"
    [] keyword != ved →
        "définir bal"
        ; attr_type := "définir type en fonction de bal"
        ; "baliser <etymon> avec attribut type = attr_type"
    FI
    ; keyword := "mot-clé suivant détecté dans vs"
OD
]|

partition-lemmes()
    balises terminales: p
    balises skipped: b, i, sc, e, lang_etymon
    balises invisibles: appelnote
    balises visibles: par défaut
    balises inattendues: entry

partition-sous-lemmes()
    balises terminales: p
    balises skipped: b, i, sc, e, lang_etymon
    balises invisibles: appelnote
    balises visibles: par défaut
    balises inattendues: entry
```

F.14 Tag-geoling

```
|[
VAR
    ck : chunk ;
    vs : virtualString
BEGIN
; "initialiser ck à la première balise ouvrante <p> après </entry>"
; DO ck != null →
    vs := "construire chaîne virtuelle à partir de <p> selon partition"
    ; "traiter les sigles réguliers dans vs"
    ; vs := "construire chaîne virtuelle à partir de <p> selon partition"
    ; "traiter les sigles erronés dans vs"
    : "résoudre les enchâssements"
    ; "initialiser ck à la prochaine balise ouvrante <p>"
OD
END
]|

partition
    balises terminales: p
    balises visibles: aucune
```

```
balises transparentes: biblio; par défaut
balises invisibles: appelnote, def, geoling, i, etymon,
(renvoi), signature, lang
balises inattendues: entry
```

F.14.1 "traiter les sigles réguliers dans vs"

La commande `"traiter les sigles réguliers dans vs"` nécessite la déclaration de variables supplémentaires :

```
status_attr : string ;
keyword, val : string
```

```
|[
"chercher dans vs tous les mots-clés de few-geoling-base"
; keyword := "premier mot-clé trouvé dans vs"
; status_attr := "ok"
; DO keyword != null →
    val := "extraire la valeur correspondant à keyword dans
    la 2e colonne de few-geoling-base"
    ; IF val = "g" →
        "baliser keyword commme <geoling> avec un attribut status
        prenant la valeur de la variable status_attr"
    [] val = "l" →
        "baliser keyword en tant que <lang> avec un attribut status prenant
        la valeur de la variable status_attr"
    FI
    ; keyword := "mot-clé suivant trouvé dans vs"
OD
]|
```

F.14.2 "traiter sigles erronés dans vs"

La commande `"traiter sigles erronés dans vs"` nécessite la déclaration d'une variable supplémentaire :

```
k_error : string
```

```
|[
"chercher dans vs tous les mots-clés de few-geoling-error-base"
; k_error := "premier mot-clé trouvé dans vs"
; DO k_error != null →
    IF "k_error se trouve dans un élément <biblio>" →
        SKIP
    [] "k_error ne se trouve pas dans un élément <biblio>" →
        keyword := "extraire le mot-clé correspondant à k_error dans
        few-geoling-error-base"
        ; "vérifier que keyword appartient à few-geoling-base"
```

```
        ; "remplacer k_error par keyword"
    FI
    ; k_error := "mot-clé suivant trouvé dans vs"
OD
; "chercher dans vs tous les mots-clés de few-geoling-base"
; keyword := "premier mot-clé trouvé dans vs"
; status_attr := "fixed"
; DO keyword != null →
    val := "extraire la valeur correspondant à keyword dans
    la 2e colonne de few-geoling-base"
    ; IF val = "g" →
        "baliser keyword en tant que <geoling> avec un attribut status
        prenant la valeur de la variable status_attr"
    [] val = "l" →
        "baliser keyword en tant que <lang> avec un attribut status prenant
        la valeur de la variable status_attr"
    FI
    ; keyword := "mot-clé suivant trouvé dans vs"
OD
]|
```

F.14.3 "résoudre les enchâssements"

La commande "`résoudre les enchâssements`" nécessite la déclaration de variables supplémentaires :

```
    ck_geo : chunk
    bal_licites : set of tags
```

```
|[
bal_licites := [<geoling>, <lang>]
; "initialiser ck_geo à la première balise de <p> appartenant à
bal_licites"
; DO ck_geo != null →
    IF "ck_geo est inclus dans un <biblio>" →
        "supprimer les balises <geoling> et </geoling> (ou <lang> et </lang>)"
        ; bib_attr := "extraire la valeur de l'attribut status de <biblio>"
        ; IF bib_attr = "geoling?" →
            "supprimer l'attribut status de <biblio>"
        [] bib_attr != "geoling?" →
            SKIP
        FI
    [] "ck_geo et son correspondant fermant incluent un <biblio>" →
        bib_attr := "extraire la valeur de l'attribut status de <biblio>"
        IF bib_attr = "geoling?" →
            "supprimer les balises <geoling> et </geoling> (ou <lang> et </lang>)"
        [] bib_attr != "geoling?" →
            "supprimer les balises <biblio> et </biblio>"
        FI
```

```
    [] ("ck_geo n'est pas inclus dans un <biblio>") AND
       ("ck_geo et son correspondant fermant n'incluent pas un <biblio>") →
        SKIP
    FI
    ; "déplacer ck_geo à la prochaine balise de <p> appartenant à bal_licites"
OD
]|
```

F.15 Tag-gram

```
|[
VAR
    ck : chunk ;
    vs : virtualString
BEGIN
"initialiser ck à la balise ouvrante <entry>"
; vs := "construire chaîne virtuelle à partir de <entry>
         selon partition-entry()"
; "traiter vs"
; "déplacer ck à la première balise ouvrante <p> après </entry>"
; DO ck != null →
    vs := "construire chaîne virtuelle à partir de <p> selon partition-p()"
    ; "traiter vs"
    ; "déplacer ck à la prochaine balise ouvrante <p>"
OD
END
]|

partition-entry()
    balises terminales: entry
    balises visibles: aucune
    balises transparentes: par défaut
    balises invisibles: e, def, date, geoling, i, lang_etymon, pnum, pref, renvoi
    balises inattendues: p

partition-p()
    balises terminales: p
    balises visibles: aucune
    balises transparentes: par défaut
    balises invisibles: e, def, date, geoling, i, lang_etymon, pnum, pref, renvoi
    balises inattendues: entry
```

F.15.1 "traiter vs"

La commande "`traiter vs`" nécessite la déclaration d'une variable supplémentaire :

```
    keyword : string
```

```
|[
"chercher tous les mots-clés de few-gram-base dans vs"
; keyword := "premier mot-clé trouvé dans vs"
; DO keyword != null →
    "baliser keyword en tant que <gram>"
    ; keyword := "prochain mot-clé dans vs"
OD
]|
```

F.16 Tag-form

```
|[
VAR
    ck, ck_i : chunk ;
    vs : virtualString
BEGIN
"initialiser ck à la première balise ouvrante <p> après </entry>"
; DO ck != null →
    vs := "construire une chaîne virtuelle à partir de <p> selon partition()"
        //=> hors <entry>, <def>, <etymon>, <pnum>, <pref>
    ; "initialiser ck_i à la première balise ouvrante <i> dans vs"
    ; DO ck_i != null →
        "traiter une section en italiques dans vs"
        ; "déplacer ck_i à la prochaine balise ouvrante <i> dans vs"
    OD
    ; "déplacer ck à la prochaine balise ouvrante <p>"
OD
END
]|

partition()
    balises terminales: p
    balises visibles: i
    balises transparentes: par défaut
    balises invisibles: def, etymon, pnum, pref, e, appelnote, affix
    balises inattendues: aucune
```

F.16.1 "traiter une section en italiques dans vs"

La commande `"traiter une section en italiques dans vs"` nécessite la déclaration de variables supplémentaires :

```
    num_car : integer ;
    num_car_sans_espace : integer
```

```
|[
num_car := "calculer nombre de caractères entre <i> et </i> dans vs"
```

```
; num_car_sans_espace := "num_car diminué du nombre de caractères d'espacement"
; IF num_car_sans_espace = 0 →
    SKIP //ne pas baliser
[] (num_car_sans_espace = 1) AND ("c'est un caractère grec") →
    SKIP //ne pas baliser
[] ((num_car_sans_espace = 1) AND ("ce n'est pas un caractère grec"))
    OR (num_car_sans_espace > 1) →
    "définir les attributs"
    ; IF "<p> est inclus dans un élément <doc>" →
        "étendre <form> dans <doc>"
        ; "baliser en intégrant l'extension"
    [] "<p> n'est pas inclus dans un élément <doc>" →
        "baliser la section en italiques en ajoutant les attributs (si définis)"
    FI
FI
]|
```

F.16.2 "définir les attributs"

La commande "`définir les attributs`" nécessite la déclaration de variables supplémentaires :

```
    car_grec : boolean ;
    attr_lang, attr_type : string ;
    mr_tiret : string ;
    pat_tiret : regex ;
    forme_composee : boolean

|[
car_grec := "calculer s'il y a au moins deux caractères grecs consécutifs dans vs"
; pat_tiret := /^(\s?\-[^\-]{1,3}\s?)|(\s?[^\-]{1,3}\-\s?)/
              { e.g. "a-", "-et", "-abc" mais pas "abcd-" }
; mr_tiret := "matcher pat_tiret dans vs"
; "identifier forme composée de plusieurs mots"
   { e.g. "il était à notre ouache", "dock flottant" mais pas "wagon-salon",
   "wagon" }
; IF (num_car_sans_espace = 1) →
    attr_type := "?"
    ; "émettre un avertissement"
[] ((num_car_sans_espace > 1) AND (car_grec = true)) →
    attr_lang := "greek"
[] ((num_car_sans_espace > 1) AND (car_grec = false) AND
    (mr_tiret != null)) →
    attr_type := "?"
    ; "émettre un avertissement"
[] ((num_car_sans_espace > 1) AND (car_grec = false) AND
    (mr_tiret = null) AND (forme_composee = true)) →
    IF "<p> est inclus dans <com> ou dans <notes>" →
        attr_type := "citation?"
```

```
    [] "<p> n'est pas inclus dans <com> ou dans <notes>" →
    attr_type := "locution?"
    FI
[] ((num_car_sans_espace > 1) AND (car_grec = false) AND
    (mr_tiret = null) AND (forme_composee = false)) →
    SKIP { pas d'attributs }
FI
]|
```

F.16.3 "identifier forme composée de plusieurs mots"

La commande "identifier forme composée de plusieurs mots" nécessite la déclaration de variables supplémentaires :

```
  word_count : integer ;
  offset : integer ;
  len : integer ;
  c : char ;
  form_word_char_count : integer

|[
word_count = 0
; offset := 0
; len := "nombre de caractères de vs"
; DO (offset > -1) AND (offset < len) →
    c := "caractère de vs à l'index offset"
    ; form_word_char_count := 0
    ; DO (c != null) AND (offset < len) →
        IF "c n'est pas un nombre ou un délimiteur" →
            form_word_char_count := form_word_char_count + 1
            ; offset := offset + 1
            ; c := "caractère suivant dans vs"
        [] "c est un nombre ou un délimiteur" →
            c := null
        FI
    OD
    ; IF form_word_char_count < 2 →
        offset := -1
    [] form_word_char_count >= 2 →
        IF (offset < len) →
            IF "le caractère de vs à l'index offset n'est pas un espace" →
                offset := -1
            [] "le caractère de vs à l'index offset est un espace" →
                offset := offset + 1
                word_count := word_count + 1
            FI
        [] offset = len →
            word_count := word_count + 1
        FI
```

```
    FI
OD
; forme_composee := (word_count >= 2)
]|
```

F.16.4 "étendre <form> dans <doc>"

La commande "`étendre <form> dans <doc>`" nécessite la déclaration de variables supplémentaires :

```
    extended_to_left : boolean ;
    extended_to_right : boolean ;
    left_tag : chunk ;
    vs_left, vs_right : virtualString ;
    pat_left, pat_right : regex ;
    mr : string

|[
extended_to_left := false
; extended_to_right := false
; "initialiser left_tag à la première balise à gauche de <i>"
; DO (left_tag = </e>) OR
     (left_tag = </appelnote>) OR
     (left_tag = <lb/>) OR
     (left_tag = <col/>) →
    "déplacer left_tag à la prochaine balise à gauche"
OD
; vs_left := "construire chaîne virtuelle à partir de left_tag jusque <i>
selon partition-gauche()"
; pat_left := /([A-Za-z]+\s*)+/
; mr := "matcher pat_left dans vs_left en activant l'aplatissement des
caractères"
; IF mr != null →
    extended_to_left := true
    ; "inclure mr dans la section à baliser"
[] mr = null →
    SKIP
FI
; vs_right := "construire chaîne virtuelle à partir de </i>
selon partition-droite()"
; pat_right := /(\s*[A-Za-z]+)+/
; mr := "chercher pat_right dans vs_right en activant l'aplatissement des
caractères"
; IF mr != null →
    extended_to_right := true
    ; "inclure mr dans la section à baliser"
[] mr = null →
    SKIP
FI
```

```
; IF (extended_to_left = true) OR (extended_to_right = true) →
    "émettre un avertissement"
[] (extended_to_left = false) AND (extended_to_right = false) →
    SKIP
FI
]|

partition-gauche()
    balises terminales: i; par défaut
    balises visibles: aucune
    balises transparentes: lb, col
    balises invisibles: e, appelnote
    balises inattendues: biblio, def, form, geoling, gram, p

partition-droite()
    balises terminales: p, biblio, def, form, geoling, gram; par défaut
    balises visibles: aucune
    balises transparentes: lb, col
    balises invisibles: e, appelnote
    balises inattendues: aucune
```

F.17 Tag-lang-etymon

```
|[
VAR
    vs : virtualString ;
    lang : string ;
    ck : chunk
BEGIN
"initialiser ck à la balise ouvrante <entry>"
; vs := "construire chaîne virtuelle à partir de <entry> selon partition()"
; lang := ""
; "traiter vs"
; "attribuer une langue à l'étymon-vedette"
END
]|

partition()
    balises terminales = entry
    balises visibles = etymon, b
    balises transparentes = par défaut
    balises invisibles = appelnote, e
    balises inattendues = aucune
```

F.17.1 "traiter vs"

La commande "traiter vs" nécessite la déclaration d'une variable supplémentaire :

```
    mr : string

|[
mr := "rechercher tous les mots-clés de few-lang-etymon-base dans vs"
; IF mr = null →
    SKIP
[] mr != null →
    IF "plusieurs mots-clés trouvés" →
        "alerte et arrêt du traitement"
    [] "un seul mot-clé trouvé" →
        "insérer, juste avant mr, une balise <lang-etymon>"
        ; "insérer, juste après mr, une balise </lang-etymon>"
        ; lang := "mr sans les parenthèses"
    FI
FI
]|
```

F.17.2 "attribuer une langue à l'étymon-vedette"

```
|[
"déplacer ck à la première balise <etymon> de vs dont l'attribut type a la
valeur "vedette""
; IF lang != "" →
    SKIP
[] lang = "" →
    "chercher le contenu de <etymon> dans few-onomatop-base"
    ; IF "test positif" →
        lang := "onomatopee"
    [] "test négatif" →
        "vérifier si l'étymon commence par une majuscule"
        ; IF "test positif" →
            lang := "nom_propre"
        [] "test négatif" →
            "définir une langue implicite"
        FI
    FI
FI
; "ajouter à la balise <etymon> un attribut lang prenant
   la valeur de la variable lang"
]|
```

F.17.3 "définir une langue implicite"

La commande "`définir une langue implicite`" nécessite la déclaration d'une variable supplémentaire :

```
    vol_id : integer
```

```
|[
vol_id := "extraire le numéro de volume de l'article"
; IF ((vol_id >= 1) AND (vol_id <= 14)) OR
    (vol_id = 24) OR
    (vol_id = 25) →
    lang := "lt."
[] vol_id = 18 →
    lang := "e."
[] vol_id = 20 →
    "définir une langue pour le volume 20"
[] ((vol_id >= 15) AND (vol_id <= 17)) OR
    (vol_id = 19) OR
    ((vol_id >= 21) AND <vol_id <= 23)) →
    SKIP
FI
]|
```

F.17.4 "définir une langue pour le volume 20"

La commande `"définir une langue pour le volume 20"` nécessite la déclaration d'une variable supplémentaire :

```
    pg_id : integer
```

```
pg_id := "extraire le numéro de page de l'article"
; IF ((pg_id >= 1) AND (pg_id < 18)) OR
    (pg_id = 116) →
    lang := "bret."
[] pg_id = 18 →
    lang := "bask."
[] (pg_id >= 24) AND (pg_id <= 28) →
    lang := "hebr."
[] (pg_id = 29) OR (pg_id = 30) →
    lang := "zig."
[] (pg_id = 31) OR (pg_id = 32) →
    lang := "magy."
[] pg_id = 54 →
    lang := "eskimo"
[] pg_id = 114 →
    lang := "austr."
[] ((pg_id > 18) AND (pg_id < 24)) OR
    ((pg_id > 32) AND (pg_id < 54)) OR
    ((pg_id > 54) AND (pg_id < 114)) OR
    (pg_id = 115) OR
    (pg_id > 116) →
    SKIP
FI
```

F.18 Tag-microstructure

```
|[
VAR
    prev_id : string ;
    ck : chunk ;
    p_licites : set of tags
BEGIN
prev_id := "unknown"
; p_licites := [<doc>, <mixt>]
; "initialiser ck à la première balise de l'article
   appartenant à p_licites"
; DO ck != null →
    "identifier le paragraphe"
    ; "détecter les titres"
    ; IF "au moins un titre a été détecté" →
        "baliser les groupes"
    [] "aucun titre n'a été détecté" →
        SKIP
    FI
    ; "déplacer ck à la prochaine balise appartenant à p_licites"
OD
END
]|
```

F.18.1 "identifier le paragraphe"

La commande "`identifier le paragraphe`" nécessite la déclaration d'une variable supplémentaire :

```
    ck_p : chunk
```

```
|[
"déplacer ck_p à la première balise ouvrante <pnum> après ck"
; IF ck_p != null →
    prev_id := "extraire de <pnum> la valeur de l'attribut id"
[] ck_p = null →
    SKIP
FI
; "ajouter à <p> un attribut structid prenant la valeur
   de la variable prev_id"
]|
```

F.18.2 "détecter les titres"

La commande "`détecter les titres`" nécessite la déclaration d'une variable supplémentaire :

```
    tiret : boolean

|[
"déplacer ck_p à la première balise ouvrante <title> après ck"
; DO ck_p != null →
    tiret := false
    ; "chercher un tiret cadratin devant le titre"
    ; IF tiret = true →
        "mémoriser l'emplacement de <title> dans une liste"
    [] tiret = false →
        SKIP
    FI
    ; "déplacer ck_p à la prochaine balise ouvrante <title> du paragraphe"
OD
]|
```

F.18.3 "chercher un tiret cadratin devant le titre"

La commande "chercher un tiret cadratin devant le titre" nécessite la déclaration de variables supplémentaires :

```
    vs : virtualString
    continue : boolean
    c : char

|[
vs := "construire une chaîne virtuelle selon partition()"
; continue := true
; "initialiser c au premier caractère à gauche de la balise <title>"
; DO continue = true →
    IF (c = '-') OR (c = '--') →
        continue := false
        ; tiret := true
    [] c = ' ' →
        SKIP //continue = true
    [] (c != '-') AND (c != '--') AND (c != ' ') →
        continue := false
        ; tiret := false
    FI
    ; "déplacer c au caractère précédent à gauche"
OD
]|

partition()
    balises terminales: title
    balises visibles: aucune
    balises transparentes: par défaut
    balises invisibles: affix, appelnote, b
    balises inattendues: p
```

F.18.4 "baliser les groupes"

```
|[
"déplacer ck_p à la première balise <title> mémorisée dans la liste"
"insérer, juste devant <title>, une balise <group>"
; "déplacer ck_p à la prochaine balise <title> mémorisée dans la liste"
; DO ck_p != null →
    "insérer </group><group> juste devant <title>"
    ; "déplacer le tiret cadratin à droite de </group>"
    ; "déplacer ck_p à la prochaine balise <title> mémorisée dans la liste"
OD
; "insérer </group> devant </p>"
]|
```

F.19 Tag-notes

```
|[
VAR
    n : integer ;
    ck : chunk ;
    vs : virtualString
BEGIN
"initialiser ck à la balise ouvrante <notes>"
; IF ck = null →
    SKIP
[] ck != null →
    n := 0
    ; "déplacer ck à la première balise ouvrante <p> après <notes>"
    ; DO ck != null →
        n := n + 1
        vs := "construire chaîne virtuelle à partir de <p>
               selon partition()"
        ; "traiter vs"
        ; "déplacer ck à la balise ouvrante <p> suivante"
    OD
    ; "déplacer ck à la balise ouvrante <notes>"
    ; "ajouter à <notes> un attribut count avec n pour valeur"
FI
END
]|

partition()
    balises terminales: p
    balises visibles: aucune
    balises transparentes: par défaut
    balises invisibles: aucune
    balises inattendues: aucune
```

F.19.1 "traiter vs"

La commande "`traiter vs`" nécessite la déclaration de variables supplémentaires :

```
    mr : string ;
    num : integer ;
    pat : regex

|[
pat := /\s*([1-9][0-9]*)\)\s*/
; mr := "chercher pat dans vs"
; IF mr = null →
    "alerte et arrêt du traitement"
[] mr != null →
    SKIP
FI
; IF "il y a des caractères entre <p> et mr" →
    "alerte et arrêt du traitement"
[] "il n'y a pas de caractères entre <p> et mr" →
    SKIP
FI
; num := "extraire de mr le numéro de note"
; IF num != n →
    "alerte et arrêt du traitement"
[] num = n →
    SKIP
FI
; "insérer, juste après <p>, une balise <note> avec un
attribut id prenant la valeur de la variable num"
; "insérer </note> juste avant le </p> suivant"
; "déplacer ck juste après la balise </note> qui vient d'être insérée"
]|
```

F.20 Tag-numbering

```
|[
VAR
  ck : chunk ;
  type_attr : string ;
  pnum_count : integer ;
  numbering_store : set of string

BEGIN
  "initialiser ck à la balise ouvrante <art>"
  ; type_attr := "extraire de <art> la valeur de l'attribut type"
  ; IF type_attr = "renvoi" →
      SKIP
    [] type_attr != "renvoi" →
```

```
       pnum_count := 0
       ; numbering_store := [ ]
       ; "détecter les repères de numérotation"
       ; IF pnum_count > 0 →
            "détecter les références de numérotation"
            ; "calculer le cross-coverage"
         [] pnum_count = 0 →
            SKIP
         FI
    FI
END
]|
```

F.20.1 "détecter les repères de numérotation"

La commande `"détecter les repères de numérotation"` nécessite la déclaration de variables supplémentaires :

```
  bal_licites : set of tags ;
  vs : virtualstring

|[
  bal_licites := [<doc>, <mixt>]
  ; "initialiser ck à la balise <p> incluse dans la première balise
     du chunk stream appartenant à bal_licites"
  ; DO ck != null →
       vs := "construire chaîne virtuelle à partir de <p> selon
              partition-pnum()"
       ; "détecter le repère de numérotation du paragraphe"
       ; "déplacer ck à la balise <p> incluse dans la balise suivante
          du chunk stream appartenant à bal_licites"
    OD
]|

partition-pnum()
  balises terminales: p
  balises visibles: aucune
  balises transparentes: par défaut
  balises invisibles: appelnote, e
  balises inattendues: aucune
```

F.20.2 "détecter le repère de numérotation du paragraphe"

La commande `"détecter le repère de numérotation du paragraphe"` nécessite la déclaration de variables supplémentaires :

```
  tok : string ;
  numbering_item : string
```

```
|[
  "tokeniser vs en utilisant le point et l'espace comme séparateurs"
  ; "initialiser tok au premier token"
  ; numbering_item := ""
  ; DO tok != null →
       IF "marque de numérotation identifiée dans tok" →
          "concaténer tok, ainsi que l'éventuel séparateur le
           précédant, à numbering_item"
          ; "assigner comme valeur à tok le token suivant"
       [] "pas de marque de numérotation identifiée dans tok" →
          tok := null
       FI
    OD
  ; IF (numbering_item = "") = false →
       pnum_count := pnum_count + 1
       ; "baliser en tant que <pnum> la section de vs
          correspondant à numbering_item"
       ; "ajouter à la balise <pnum> un attribut id dont la valeur est
          la version décodée de numbering_item"
       ; "ajouter à numbering_store la version décodée de numbering_item"
       ; "désambiguiser la balise <pnum>, en supprimant les éventuelles
          balises <geoling>, <lang> et <biblio>, incluses et englobantes"
    [] numbering_item = "" →
       SKIP
    FI
]|
```

F.20.3 "détecter les références de numérotation"

La commande `"détecter les références de numérotation"` nécessite la déclaration d'une variable supplémentaire :

```
  vs_i : virtualstring
```

```
|[
  "initialiser ck à la balise <p> incluse dans la première balise <com>"
  ; DO ck != null →
       vs_i := "construire chaîne virtuelle à partir de <p>
                selon partition-pref()"
       ; DO vs_i != null →
            ; "détecter les références de numérotation du paragraphe"
            vs_i := "construire chaîne virtuelle juste après vs_i
            selon partition-pref()"
        OD
       ; "déplacer ck à la balise <p> incluse dans la balise <com> suivante"
    OD
]|
```

```
partition-pref()
```

```
balises terminales: p
balises visibles: renvoi
balises transparentes: par défaut
balises invisibles: appelnote, e, biblio, def
balises inattendues: aucune
```

F.20.4 "détecter les références de numérotation du paragraphe"

La commande "détecter les références de numérotation du paragraphe" nécessite la déclaration de variables supplémentaires :

```
tok : string ;
numbering_item : string
```

```
|[
  "tokeniser vs_i (1) en utilisant comme séparateurs l'ensemble des marques
   de numérotation dérivé des repères de numérotation précédemment détectés,
   (2) en vérifiant la présence de texte entre les séparateurs consécutifs,
   et (3) en produisant explicitement un token pour chaque séparateur"
  ; "initialiser tok au premier token"
  ; numbering_item := ""
  ; DO tok != null →
       IF "tok est une marque de numérotation" →
          IF "marque de numérotation inclusable
              dans une référence de numérotation" →
             IF "marque de numérotation est la première d'une référence
                 de numérotation" →
                "traiter numbering_item, puis le réinitialiser"
             [] "marque de numérotation n'est pas la première d'une
                 référence de numérotation" →
                SKIP
             FI
             ; "concaténer tok, précédé d'un espace s'il n'est le premier
                de la référence de numérotation, à numbering_item"
          [] "marque de numérotation non inclusable" →
             "traiter numbering_item, puis le réinitialiser"
          FI
       [] "tok n'est pas une marque de numérotation" →
          SKIP
       FI
       ; "assigner comme valeur à tok le token suivant"
    OD
  ; "traiter numbering_item, puis le réinitialiser"
]|
```

F.20.5 "traiter numbering_item, puis le réinitialiser"

```
|[
```

```
  IF (numbering_item = "") = false →
     ; "décoder l'implicite de la référence de numérotation incluse
        dans numbering_item,
        en fonction (en ordre inverse) des références de numérotation
        déjà balisées et de numbering_store"
     ; "baliser en tant que <pref> la section de vs correspondant
        à numbering_item"
     ; "ajouter à la balise <pref> un attribut id dont la valeur est la
        version décodée de numbering_item"
     ; "désambiguiser la balise <pref>, en supprimant les éventuelles
        balises <geoling>, <lang> et <biblio>, incluses et englobantes"
     ; numbering_item = ""
  [] numbering_item = "" →
     SKIP
  FI
]|
```

F.21 Tag-precisions

```
VAR
    ck : chunk ;
    p_licites : set of tags
BEGIN
p_licites := [<mixt>, <doc>]
; "initialiser ck à la première balise appartenant à p_licites"
; DO ck != null →
    "déplacer ck à la balise ouvrante <p> située à sa droite"
    ; "traiter les précisions entre parenthèses"
    ; "traiter les éléments de précision hors parenthèses"
    ; "supprimer les faux <gram>"
    ; "déplacer ck à la prochaine balise ouvrante appartenant à p_licites"
OD
END
```

F.21.1 "traiter les précisions entre parenthèses"

La commande `"traiter les précisions entre parenthèses"` nécessite la déclaration de variables supplémentaires :

```
     vs : virtualString ;
     paire : string
```

```
vs := "construire chaîne virtuelle à partir de <p> selon partition()"
; "vérifier la bonne imbrication des parenthèses"
; paire := "identifier dans vs le premier extrait
    situé entre une parenthèse ouvrante et une parenthèse fermante
    non englobées dans d'autres parenthèses"
```

```
; DO paire != null →
    "insérer <precisions> avant la parenthèse ouvrante"
    ; "insérer </precisions> après la parenthèse fermante"
    ; "définir la valeur de l'attribut status"
    ; paire := "identifier dans vs le prochain extrait
                situé entre une parenthèse ouvrante et une parenthèse fermante
                non englobées dans d'autres parenthèses"
OD

partition()
    balises terminales: p
    balises visibles: aucune
    balises transparentes: par défaut
    balises invisibles avec contenu: appelnote, def, etymon, form, pnum
    balises inattendues: aucune
```

F.21.2 "définir la valeur de l'attribut status"

La commande `"définir la valeur de l'attribut status"` nécessite la déclaration de variables supplémentaires :

```
    contains_biblio, contains_date : boolean ;
    contains_geoling, contains_form : boolean ;
    status_attr : string
```

```
contains_biblio := "déterminer si <precisions> contient au moins un
élément <biblio>"
; contains_date := "déterminer si <precisions> contient au moins un
  élément <date>"
; contains_geoling := "déterminer si <precisions> contient au moins
  un élément <geoling>"
; contains_form := "déterminer si <precisions> contient au moins un
  élément <form>"
; IF (contains_geoling = true) AND (contains_form = true) →
    status_attr := "contains-unit"
[] ((contains_geoling = false) OR (contains_form = false)) AND
   ((contains_biblio = true) OR (contains_date = true)) →
    status_attr := "ok"
[] ((contains_geoling = false) OR (contains_form = false)) AND
   (contains_biblio = false) AND (contains_date = false) →
    status_attr := "ambigu"
FI
; "ajouter à <precisions> un attribut status prenant la valeur de
la variable status_attr"
```

F.21.3 "traiter les éléments de précision hors parenthèses"

La commande `"traiter les éléments de précision hors parenthèses"` nécessite la déclaration de variables supplémentaires :

```
    bal : tag

bal := <biblio>
; "traiter vs"
; bal := <date>
; "traiter vs"
```

F.21.4 "traiter vs"

La commande "`traiter vs`" nécessite la déclaration de variables supplémentaires :

```
    texte_licite : boolean ;
    ck_bal : chunk ;
    ck_prev : chunk

"initialiser ck_bal à la première balise du paragraphe correspondant
 à la valeur de la variable bal qui n'est pas dans un élément <precisions>"
; DO ck_bal != null →
        "initialiser ck_prev à la balise fermante la plus proche à gauche
         de ck_bal"
        ; IF (ck_prev = </def>) OR
             (ck_prev = </form>) OR
             (ck_prev = </geoling>) OR
             (ck_prev = </gram>) →
            vs := "construire chaîne virtuelle à partir de ck_prev selon
                   partition-vs(ck_bal)"
            ; texte_licite := "vérifier le texte entre ck_prev et ck_bal"
            ; IF texte_licite = true →
                    "baliser ck_bal, la balise fermante correspondante et leur contenu
                    en tant que <precisions>"
            [] texte_licite = false →
                SKIP
            FI
        [] (ck_prev != </def>) AND
             (ck_prev != </form>) AND
             (ck_prev != </geoling>) AND
             (ck_prev != </gram>) →
            SKIP
        FI
    ; "déplacer ck_bal à la prochaine balise du paragraphe correspondant à la
       valeur de la variable bal qui n'est pas dans un élément <precisions>"
OD

partition-vs(tag ck_bal)
    balises terminales: ck_bal
    balises visibles: par défaut
    balises transparentes: lb, col
    balises invisibles: appelnote, def, etymon, form
    balises inattendues: aucune
```

F.21.5 "vérifier le texte entre prev-tag et tag"

La commande "`traiter vs`" nécessite la déclaration de variables supplémentaires :

```
    pat1, pat2 : regex ;
    mr1, mr2 : string
```

```
pat1 := /[.;----]/
; pat2 := /id\./
; mr1 := "matcher pat1 dans vs"
; IF mr1 = null →
    texte_licite := true
[] mr1 != null →
    mr2 := "matcher pat2 dans vs"
    ; IF mr2 != null →
        texte_licite := true
    [] mr2 = null →
        texte_licite := false
    FI
FI
```

F.21.6 "supprimer les faux <gram>"

```
"déplacer ck à la première balise <precisions> après <p>"
: DO ck != null →
    "déplacer ck_bal à la première balise <gram> incluse
    dans l'élément <precisions> débutant à ck"
    ; DO ck_bal != null →
        bal := "chercher, avant </precisions>, la balise ouvrante qui suit
                la balise fermante </gram> correspondant à ck_bal"
        ; IF bal = <biblio> →
            "traiter gram"
        [] bal != <biblio> →
            SKIP
        FI
        ; "déplacer ck_bal à la prochaine balise <gram> incluse
           dans l'élément <precisions> débutant à ck"
    OD
    ; "déplacer ck à la prochaine balise <precisions> avant </p>"
OD
```

F.21.7 "traiter gram"

La commande "`traiter gram`" nécessite la déclaration de variables supplémentaires :

```
    mr : string ;
    pat : regex
```

```
vs := "construire vs à partir de <gram> selon partition-gram()"
; pat := /v\.\s*/
; mr := "chercher pat dans vs"
; IF mr != null →
    "supprimer <gram> et </gram> autour de mr"
[] mr = null →
    SKIP
FI

partition-gram()
    balises terminales: gram
    balises visibles: aucune
    balises transparentes: par défaut
    balises invisibles: aucune
    balises inattendues: aucune
```

F.22 Tag-renvoi

```
|[
BEGIN
"traiter l'entrée"
; "traiter les renvois avec étymon"
; "traiter les renvois sans étmon"
END
]|
```

F.22.1 "traiter l'entrée"

La commande `"traiter l'entrée"` nécessite la déclaration de variables supplémentaires :

```
    ck : chunk ;
    art_renvoi : boolean ;
    keyword : string
```

```
|[
"initialiser ck à la balise ouvrante <entry>"
; "déplacer ck à la première balise <etymon> après <entry>"
; keyword := "extraire le contenu de l'élément <etymon>"
; IF "keyword appartient à few-renvoi-base" →
    "vérifier qu'il s'agit d'un article de renvoi et le baliser"
[] "keyword n'appartient pas à few-renvoi-base" →
    art_renvoi := false
FI
]|
```

F.22.2 "vérifier qu'il s'agit d'un article de renvoi et le baliser"

La commande "`vérifier qu'il s'agit d'un article de renvoi et le baliser`" nécessite la déclaration de variables supplémentaires :

```
    ck_e : chunk ;
    np, ne : integer

|[
np := "calculer le nombre de paragraphes de l'article (hors entrée)"
; IF np > 1 →
    art_renvoi := false
[] np = 1 →
    "vérifier que le paragraphe commence par S. et contient un étymon"
    ; IF "le test est positif" →
        art_renvoi := true
        ; "baliser tout le contenu du paragraphe en tant que renvoi"
    [] "le test est négatif" →
        art_renvoi := false
    FI
[] np = 0 →
    ne := "calculer le nombre d'étymon dans l'entrée"
    ; IF ne = 1 →
        "alerte et arrêt du traitement"
    [] ne = 2 →
        "baliser le deuxième étymon en tant que renvoi"
        ; art_renvoi := true
    [] ne > 2 OR ne = 0 →
        "alerte et arrêt du traitement"
    FI
FI
]|
```

F.22.3 "traiter les renvois avec étymon"

La commande "`traiter les renvois avec étymon`" nécessite la déclaration de variables supplémentaires :

```
    ck_e : chunk ;
    mr : string ;
    vs_left, vs_right : virtualString

|[
IF art_renvoi = true →
    SKIP
[] art_renvoi = false →
    "définir les expressions régulières regexRenvoiAvantEtymon,
    regexExtensionEtymon, regexRenvoiApresEtymon,
    regexRenvoiSansEtymon, regexRenvoiMultiple"
```

```
    ; "déplacer ck à la première balise ouvrante <p> après </entry>"
    ; DO ck != null →
        "initialiser ck_e à la première balise ouvrante <etymon> après <p>"
        ; DO ck_e != null →
            "définir le début de vs_left"
            ; vs_left := "construire chaîne virtuelle selon partition-left"
            ; mr := "chercher regexRenvoiAvantEtymon dans vs_left"
            ; IF mr != null →
                "baliser le renvoi"
                ; mr := "chercher regexExtensionEtymon après la balise
                </renvoi> insérée"
                ; IF mr != null →
                    "baliser mr en tant que <rpref>"
                    ; "déplacer </renvoi> après mr"
                    ; "baliser les renvois multiples éventuels"
                [] mr = null →
                    SKIP
                FI
            [] mr = null →
                vs_right := "construire chaîne virtuelle selon partition-right"
                ; mr := "chercher regexRenvoiApresEtymon dans vs_right"
                ; IF mr != null →
                    "baliser le renvoi depuis ck_e jusque mr inclus"
                    ; "baliser les renvois multiples éventuels"
                [] mr = null →
                    SKIP
                FI
            FI
            ; "déplacer ck_e à la prochaine balise <etymon> non inclue
            dans un élément <renvoi>"
        OD
        ; "déplacer ck à la prochaine balise <p>"
    OD
FI
]|

partition-left()
    balises terminales : etymon
    balises visibles : aucune
    balises transparentes : par défaut
    balises invisibles : aucune
    balises inattendues : aucune

partition-right()
    balises terminales : p, etymon
    balises visibles : aucune
    balises transparentes : b, e, i, sc, par défaut
    balises invisibles : lang_etymon
    balises inattendues : aucune
```

F.22.4 "définir le début de vs_left"

```
|[
IF "il y a une balise </renvoi> à gauche de ck_e" →
    "vs_left commence à </renvoi>"
[] "il n'y a pas de balise </renvoi> à gauche de ck_e" →
    IF "il y a une balise </etymon> à gauche de ck_e" →
        "vs_left commence à </etymon>"
    [] "il n'y a pas de balise </etymon> à gauche de ck_e" →
        "vs_left commence à <p>"
    FI
FI
]|
```

F.22.5 "traiter les renvois sans étymon"

La commande `"traiter les renvois sans étymon"` nécessite la déclaration de variables supplémentaires :

```
    vs : virtualString
```

```
|[
IF art_renvoi = true →
    SKIP
[] art_renvoi = false →
    "déplacer ck à la première balise <p> après </entry>"
    ; DO ck != null →
        vs := "construire chaîne virtuelle à partir de ck selon partition"
        ; mr := "chercher regexRenvoiSansEtymon dans vs, hors éléments
        de <renvoi>"
        ; DO mr != null →
            "baliser mr en tant que <renvoi>"
            ; "retirer de vs la section de texte balisée"
            ; mr := "chercher regexRenvoiSansEtymon dans vs, hors éléments
            de <renvoi>"
        OD
        ; ck := "déplacer ck à la prochaine balise <p> après </entry>"
    OD
FI
]|

partition()
    balises terminales : p, etymon
    balises visibles : aucune
    balises transparentes : b, e, i, sc, par défaut
    balises invisibles : lang_etymon
    balises inattendues : aucune
```

F.22.6 "baliser les renvois multiples éventuels"

La commande "`baliser les renvois multiples éventuels`" nécessite la déclaration de variables supplémentaires :

```
    mr_mult : string
```

```
|[
mr_mult := "chercher regexRenvoiMultiple juste après </renvoi>"
; DO mr_mult != null →
    "baliser mr_mult en tant que <renvoi>"
    ; mr_mult := "chercher regexRenvoiMultiple juste après la balise </renvoi>
    qui vient d'être insérée"
OD
]|
```

F.23 Tag-signature

```
|[
VAR
    type_art : string ;
    etymon_vedette : string ;
    vol_id : integer ;
    pg_id : integer ;
    attr_lang : string ;
    attr_author : string ;
    spliced_nominal_signature : boolean

BEGIN
type_art := "extraire de <art> la valeur de l'attribut type"
; IF type_art = "renvoi" →
    SKIP
[] type_art != "renvoi" →
    etymon_vedette := "extraire l'étymon-vedette dans <entry>"
    ; vol_id := "extraire de la balise <few> la valeur de l'attribut volume"
    ; pg_id := "extraire le numéro de page de la première <col/>"
    ; "baliser les signatures de l'article"
    ; IF spliced_nominal_signature = true →
        SKIP
      [] spliced_nominal_signature = false →
        "insérer signature implicite"
      FI
FI
END
]|
```

F.23.1 "baliser les signatures de l'article"

La commande "`baliser les signatures de l'article`" nécessite la déclaration de variables supplémentaires :

```
    ck : chunk ;
    vs : virtualString

|[
"initialiser ck à la première balise <p>"
; DO ck != null →
    ; vs := "construire chaîne virtuelle à partir de ck selon partition()"
    ; "baliser les signatures du paragraphe"
    ; "déplacer ck à la balise <p> suivante"
  OD
]|

partition()
    balises terminales : p
    balises visibles : aucune
    balises transparentes : par défaut
    balises invisibles : appelnote
    balises inattendues : aucune
```

F.23.2 "baliser les signatures du paragraphe"

La commande "`baliser les signatures du paragraphe`" nécessite la déclaration de variables supplémentaires :

```
    keyword : string

"chercher les mots-clés de few-signature-base dans vs"
; IF "aucun mot-clé détecté" →
    SKIP
  [] "au moins un mot-clé détecté →
    IF "il y a du texte entre le dernier mot-clé détecté et la fin du paragraphe" →
        SKIP
    [] "dernier mot-clé détecté suivi uniquement d'un point
            et de caractères d'espacement éventuels, avant la fin du paragraphe" →
        { ETAPE 1: filtrer les mots-clés }
        keyword := "dernier mot-clé détecté"
        ; DO (keyword != null) →
            IF ("keyword et l'éventuel mot-clé précédent sont séparés par /^ *[;-.] *$/") →
                keyword := "mot-clé précédent"
            [] ("keyword et l'éventuel mot-clé précédent ne sont pas séparés par / *[;-.] */") →
                "effacer les mots-clés précédents de la liste de mots-clés détectés"
                ; keyword := null
            FI
          OD
```

```
        { ETAPE 2: baliser les mots-clés filtrés }
        ; keyword := "premier mot-clé parmi les mots-clés restants"
        ; DO (keyword != null) →
            "baliser la signature"
            ; keyword := "mot-clé suivant parmi les mots-clés restants"
          OD
        { ETAPE 3: vérifier langues des signatures balisées }
        ; IF "le paragraphe courant est le dernier paragraphe principal" →
            spliced_nominal_signature := true
            ; "ajouter à la balise <art> l'attribut lang avec pour valeur
               celle de l'attribut lang de la première <signature> du paragraphe"
            IF ("les attributs lang de toutes les balises <signature>
                 insérées dans le paragraphe sont toutes identiques") →
                SKIP
            [] ("les attributs lang de toutes les balises <signature>
                 insérées dans le paragraphe ne sont pas toutes identiques") →
                "émettre une alerte oops"
            FI
          [] "le paragraphe courant n'est pas le dernier paragraphe principal" →
            SKIP
          FI
    FI
  FI
```

F.23.3 "baliser la signature"

```
"définir les attributs de la signature explicite"
; "baliser keyword en tant que <signature>
   avec des attributs author et lang prenant pour valeur, respectivement,
   celles des variables attr_author et attr_lang"
```

F.23.4 "définir les attributs de la signature explicite"

```
|[
attr_author = "auteur correspondant à keyword dans few-signature-base"
; IF (attr_author = "Hubschmid") AND (vol_id = "25") →
    attr_lang := "french"
[] attr_author = "Smiřický" →
    IF ("etymon_vedette contient un mot-clé de few-signature-smiricky-base") →
        attr_lang := "french"
    [] ("etymon_vedette ne contient pas un mot-clé de few-signature-smiricky-base") →
        attr_lang := "german"
    FI
[] (((attr_author = "Hubschmid") AND (vol_id = "25")) = false) AND
   ((attr_author = "Smiřický") = false) →
    attr_lang = "langue correspondant à keyword dans few-signature-base"
FI
]|
```

F.23.5 "insérer signature implicite"

```
"définir les attributs de la signature implicite"
; "si article sans paragraphe, ajouter un paragraphe vide avant </art>"
; "insérer <signature/> à la fin du dernier paragraphe avant le <notes> éventuel,
   avec des attributs author et lang prenant pour valeur, respectivement,
   celles des variables attr_author et attr_lang"
```

F.23.6 "définir les attributs de la signature implicite"

```
|[
IF (vol_id = 20) AND (pg_id = [33-52]) →
    attr_author := "Jänicke"
    ; attr_lang := "german"
[] (vol_id = 24) AND (etymon_vedette = "animans") →
    attr_author := "?"
    ; attr_lang := "french"
[] (((vol_id = 20) AND (pg_id = [33-52])) = false) AND
   (((vol_id = 24) AND (etymon_vedette = "animans")) = false) →
    attr_author := "Wartburg"
    ; attr_lang := "german"
FI
]|
```

F.24 Tag-title

```
|[
VAR
    ck : chunk ;
    p_licites : set of tags
BEGIN
p_licites := {<doc>, <mixt-doc>}
"initialiser ck à la première balise de l'article appartenant à p_licites"
; DO ck != null →
    "baliser les titres entre crochets"
    ; "baliser les titres en tête de paragraphe"
    ; "baliser les titres internes"
    ; "baliser les titres connus"
    ; "déplacer ck à la prochaine balise appartenant à p_licites"
OD
END
]|
```

F.24.1 "baliser les titres entre crochets"

La commande "`baliser les titres entre crochets`" nécessite la déclaration de variables supplémentaires :

```
    vs : virtualString ;
    ck_vs : chunk ;
    mr_crochet_ouvrant : char ;
    mr_crochet_fermant : char
```

```
|[
vs := "construire chaîne virtuelle à partir de <p> selon
       partition-titres-crochets()"
; DO vs != null →
    {hors <def>, <precisions>, <i>, <form>, (<geoling>, <gram>, <pnum>.
    Le texte entre crochets déjà balisé par tag-precisions n'est pas pris en
    compte}
    mr_crochet_ouvrant := "matcher /_[/ le plus à gauche dans vs"
    ; mr_crochet_fermant := "matcher /]_?/ le plus à droite dans vs"
    ; IF (mr_crochet_ouvrant != null) AND
         (mr_crochet_fermant != null) AND
         (mr_crochet_ouvrant précède mr_crochet_fermant) →
        "insérer <title> avant mr_crochet_ouvrant"
        ; "insérer </title> après mr_crochet_fermant"
    [] (mr_crochet_ouvrant != null) AND
         (mr_crochet_fermant != null) AND
         (mr_crochet_ouvrant suit mr_crochet_fermant) →
        "alerte et arrêt du traitement"
    [] ((mr_crochet_ouvrant != null) AND
         (mr_crochet_fermant = null))
         OR
         ((mr_crochet_ouvrant = null) AND
         (mr_crochet_fermant != null)) →
        "alerte et arrêt du traitement"
    [] (mr_crochet_ouvrant = null) AND
         (mr_crochet_fermant = null) →
        SKIP
    FI
    ; "construire vs à partir de la fin de la précédente vs"
OD
]|

partition-titres-crochets()
    - balises terminales: p
    - balises visibles: def, precisions, i, geoling, form,
                    gram, title, par défaut
      {ces balises doivent rester visibles pour que le crochet fermant
      soit matché dans le même chunk de texte que le crochet ouvrant.}
    - balises transparentes: lb, col, b
    - balises invisibles: e, appelnote, pnum, pref,
                    affix (pour affixes entre crochets)
    - balises inattendues: aucune
```

F.24.2 "baliser les titres en tête de paragraphe"

La commande "`baliser les titres en tête de paragraphe`" balise le texte non balisé qui se trouve au tout début d'un paragraphe, entre le début du paragraphe (après <pnum>) et la première balise <geoling>. Ce texte peut être en gras (dans un élément <b>).

La commande "`baliser les titres en tête de paragraphe`" nécessite la déclaration d'une variable supplémentaire :

```
    mr : string
```

```
"déplacer ck_vs à la balise ouvrante <p>
 ou, s'il y a un <pnum>, à la balise fermante </pnum>"
; vs := "construire chaîne virtuelle à partir de ck_vs
        selon partition-titres-debut()"
; IF vs != null →
    "insérer <title> juste avant vs"
    ; mr := "matcher /_?[.;,---]_?/ tout à la fin de vs"
    ; IF mr != null →
        "insérer </title> juste avant mr"
    [] mr = null →
        "insérer </title> juste après vs"
    FI
[] vs = null →
    SKIP
FI

partition-titres-debut()
    balises terminales: p, geoling
    balises visibles: par défaut
    balises transparentes: lb, col, b
    balises invisibles: appelnote
    balises inattendues: aucune
```

F.24.3 "baliser les titres internes"

La commande "`baliser les titres internes`" nécessite la déclaration de variables supplémentaires :

```
    ck_geo : chunk ;
    bal_sans_interet : set of tags
    bal : tag
```

```
"initialiser ck_geo à la première balise <geoling> du paragraphe"
; bal_sans_interet := {e, b, lb, col, appelnote, affix, title}
; DO ck_geo != null →
    bal := "rechercher de droite à gauche à partir de <geoling>
    la première balise fermante qui n'appartient pas à bal_sans_interet,
```

```
    ou la balise <p> si aucune autre n'est trouvée"
%   ; IF bal != null →
%       SKIP
%   [] bal = null →
%       bal := <p>
%   FI
    ; vs := "construire chaîne virtuelle à partir de bal selon
    partition-titres-internes()"
    ; mr := "matcher /_+([.,;---])[^.,;---]*/ dans vs à partir de la fin"
    ; IF mr != null →
        "insérer <title> juste après le signe de ponctuation trouvé"
        ; "insérer </title> juste avant <geoling>"
    [] mr = null →
        SKIP
    FI
    ; "déplacer ck_geo à la balise <geoling> suivante du paragraphe"
OD

partition-titres-internes()
    balises terminales: geoling
    balises visibles: par défaut
    balises transparentes: lb/, col/
    balises invisibles: e, b, appelnote, title, affix
    balises inattendues: aucune
```

F.24.4 "baliser les titres connus"

La commande "baliser les titres connus" nécessite la déclaration de variables supplémentaires :

```
    keyword : string ;
    ck_bal : chunk ;
    bal_licites : set of tags

bal_licites := [<def>, <i>, <geoling>, <lang>, <form>, <gram>, <pnum>]
; "initialiser ck_bal à la première balise du paragraphe
   appartenant à bal_licites"
; DO ck_bal != null →
    bal := "rechercher de droite à gauche à partir de ck_bal
            la première balise fermante qui n'appartient pas à bal_licites,
            ou la balise <p> si aucune autre n'est trouvée"
    ; vs := "construire chaîne virtuelle à partir de bal selon
            partition-titres-connus()"
    ; "chercher dans vs les mots-clés de la liste sélectionnée"
    ; keyword := "premier mot-clé trouvé dans vs"
    ; DO keyword != null →
        "baliser keyword en tant que <title>"
        ; mr := "matcher /_*[.;,---]/ dans vs à partir de mot-clé
                 jusque balise ouvrante visible trouvée"
```

```
        ; IF mr != null →
            "déplacer </title> juste après mr"
        [] mr = null →
            SKIP
        FI
        ; keyword := "mot-clé suivant trouvé dans vs"
    OD
    ; "déplacer ck_bal à la prochaine balise du paragraphe
       appartenant à bal_licites"
OD
```

```
partition-titres-connus()
    balises terminales: def, i, geoling, form, gram, pnum, lang
    balises visibles: aucune
    balises transparentes: par défaut
    balises invisibles: e, appelnote, precisions, title
    balises inattendues: aucune
```

F.25 Tag-unit

```
|[
VAR
    ck : chunk
BEGIN
"initialiser ck à la première balise <p> incluse dans <doc> ou <mixt>"
DO ck != null →
    "lister les molécules du paragraphe"
    ; "regrouper les molécules du paragraphe en cellules"
    ; "déplacer ck à la prochaine balise <p> incluse dans <doc> ou <mixt>"
OD
END
]|
```

F.25.1 "lister les molécules du paragraphe"

La commande `"lister les molécules du paragraphe"` nécessite la déclaration de variables supplémentaires :

```
    list_ponct : set of char ;
    list_mol : set of tag ;
    ck_t : chunk
```

```
list_ponct := [ ".", ";", ",", "-" ]
; list_mol := [ <geoling>, <form>, <gram>, <def>, <precisions> ]
; "initialiser une liste L de molécules et de ponctuation inter-molécules"
; "initialiser ck_t à la première balise ouvrante après <p>"
```

```
; DO ck_t != null →
    IF ck_t in list_mol OR ck_t in [ title ] →
        IF ck_t in list_mol  →
            IF ("ck_t inclus dans <geoling>, <form>, <gram> ou <def>") OR
               ("ck_t inclus dans <precisions>" AND (ck_t = <precisions>))  →
                alerte("les molécules récursives ne sont pas licites")
            [] ("ck_t non inclus dans <geoling>, <form>, <gram>, <def> ou <precisions>")
               ("ck_t inclus dans <precisions>" AND ((ck_t = <precisions>) = false))  →
                SKIP
            FI
        [] ck_t in [ title ] →
            IF ("ck_t inclus dans <geoling>, <form>, <gram>, <def> ou <precisions>") →
                alerte("les <title> ne peuvent être inclus dans les molécules")
            [] ("ck_t non inclus dans <geoling>, <form>, <gram>, <def> ou <precisions>")
                SKIP
            FI
        FI
        ; "ajouter ck_t dans la liste L et mémoriser son emplacement"
        ; "déplacer ck_t à la balise fermante correspondante"
        ; "chercher, à droite de ck_t et jusqu'à la première
           balise ouvrante appartenant à list_mol,
           tous les caractères appartenant à list_ponct"
        ; "ajouter à la liste L les caractères de ponctuation trouvés
           et mémoriser leur emplacement"
    [] (ck_t in list_mol = false) AND (ck_t in [ title ] = false) →
        SKIP
    FI
    ; "déplacer ck_t à la prochaine balise ouvrante
           suivant la balise fermante correspondant à la balise ouvrante traitée"
OD
```

F.25.2 "regrouper les molécules du paragraphe en cellules"

La commande `"regrouper les molécules du paragraphe en cellules"` nécessite la déclaration de variables supplémentaires :

```
VAR
    premiere_molecule_a_baliser : chunk ;
    derniere_molecule_a_baliser : chunk ;
    geoling_ck, form_ck, gram_ck, def_ck, precisions_ck : chunk

; premiere_molecule_a_baliser := null
; derniere_molecule_a_baliser := null
; geoling_ck := null
; form_ck := null
; gram_ck := null
; def_ck := null
; precisions_ck := null
; "aller au premier élément de la liste L"
```

```
; DO "il reste des éléments dans la liste L" →
    IF "l'élément courant de L est une ponctuation ou un <title>" →
        "baliser la cellule en résolvant l'implicite"
        ; premiere_molecule_a_baliser := null
        ; derniere_molecule_a_baliser := null
    [] "l'élément courant de L est une molécule de niveau égal ou,
        si la molécule n'est pas un <geoling>, inférieur
        à celui de la molécule précédente" →
        "baliser la cellule en résolvant l'implicite"
        ; premiere_molecule_a_baliser := "élement courant de la liste L"
        ; derniere_molecule_a_baliser := "élement courant de la liste L"
    [] "l'élément courant de L est une molécule de niveau supérieur
        à celui de la molécule précédente" OR
       "l'élément courant de L est le premier élément traité du paragraphe" →
        IF premiere_molecule_a_baliser = null →
            premiere_molecule_a_baliser := "élement courant de la liste L"
        [] (premiere_molecule_a_baliser = null) = false →
            SKIP
        FI
        ; derniere_molecule_a_baliser := "élement courant de la liste L"
    FI
    ; "aller à l'élément suivant dans la liste L"
OD
; IF premiere_molecule_a_baliser = null →
    {il subsiste en fin de liste L des molécules encore
     à baliser dans une cellule}
    "baliser la cellule en résolvant l'implicite"
  [] (premiere_molecule_a_baliser = null) = false →
    SKIP
  FI
```

F.25.3 "baliser la cellule en résolvant l'implicite"

La commande "baliser la cellule en résolvant l'implicite" nécessite la déclaration de variables supplémentaires :

```
VAR
    resoudre_implicite : boolean

"baliser en tant que <unit> la section du chunk stream
 délimitée par premiere_molecule_a_baliser et derniere_molecule_a_baliser"
; "itérer sur les chunks
   entre premiere_molecule_a_baliser (inclus)
   et derniere_molecule_a_baliser (inclus),
   en les assignant de manière appropriée à
   geoling_ck, form_ck, gram_ck, def_ck et precisions_ck"
; IF "élément de la liste L après derniere_molecule_a_baliser est un point ou un tiret"
     AND
     "premiere_molecule_a_baliser n'est pas une molécule <precisions>" →
```

```
    "insérer des balises <imp/> dans le chunk stream
     là où les molécules implicites de la cellule courante peuvent être explicitées
     par geoling_ck, form_ck, gram_ck, def_ck ou precisions_ck"
  [] "élément de la liste L après derniere_molecule_a_baliser
      n'est ni un point ni un tiret"
     OR
     "premiere_molecule_a_baliser est une molécule <precisions>" →
    SKIP
  FI
FI
```

Annexe G

CD-ROM

G.1 Arborescence du CD-ROM

- `articles-these-pascale-renders/` : articles FFML + version finale articles FSML
- `bases-mots-cles-these-pascale-renders/` : listes mots-clés
- `few-ra/` : logiciel complet, y compris articles, listes mots-clés, char map, guide FFML
- `introduction-cdrom-these-pascale-renders.pdf`
- `introduction-cdrom-these-pascale-renders.txt`

G.2 Schémas XML

Le schéma FFML est consultable dans le répertoire suivant :
`few-ra/src/fr/atilf/few/model/article/chunk/tag/ffml/`.

Le schéma FSML est consultable dans le répertoire suivant :
`few-ra/src/fr/atilf/few/model/article/chunk/tag/fsml/`.

G.3 Articles rétroconvertis

Les articles rétroconvertis sont consultables dans le répertoire suivante :
`articles-these-pascale-renders/`.

Table des matières

Table des figures